Artificial Intelligence Review (2025) 58:145
https://doi.org/10.1007/s10462-025-11137-6

Leveraging generative AI synthetic and social media data for content generalizability to overcome data constraints in vision deep learning

Panteha Alipour[1] · **Erika Gallegos**[1]

Accepted: 6 February 2025

Abstract
Generalizing deep learning models across diverse content types is a persistent challenge in domains like facial emotion recognition (FER), where datasets often fail to reflect the wide range of emotional responses triggered by different stimuli. This study addresses the issue of content generalizability by comparing FER model performance between models trained on video data collected in a controlled laboratory environment, data extracted from a social media platform (YouTube), and synthetic data generated using Generative Adversarial Networks. The videos focus on facial reactions to advertisements, and the integration of these different data sources seeks to address underrepresented advertisement genres, emotional reactions, and individual diversity. Our FER models leverage Convolutional Neural Networks Xception architecture, which is fine-tuned using category based sampling. This ensures training and validation data represent diverse advertisement categories, while testing data includes novel content to evaluate generalizability rigorously. Precision–recall curves and ROC-AUC metrics are used to assess performance. Results indicate a 7% improvement in accuracy and a 12% increase in precision–recall AUC when combining real-world social media and synthetic data, demonstrating reduced overfitting and enhanced content generalizability. These findings highlight the effectiveness of integrating synthetic and real-world data to build FER systems that perform reliably across more diverse and representative content.

Keywords Artificial intelligence · GenAI · Generative AI · Model generalizability · Deep learning · YouTube

✉ Erika Gallegos
erika3@colostate.edu

[1] Department of Systems Engineering, Colorado State University, 6029 Campus Delivery, Fort Collins, CO 80523, USA

Published online: 24 February 2025

1 Introduction

The generalizability of deep learning models, particularly in the domain of facial emotion recognition (FER), remains a significant challenge due to limited and non-diverse training datasets. As a result, some artificial intelligence (AI) models may induce unintended biases, such as inaccurately classifying or recognizing types of people or contexts. While FER models offer numerous benefits across diverse domains, it is important to uphold ethical responsibilities in the development and deployment of these models (Algabri et al 2024). Current and previous research has explored deep learning generalizability, yet many challenges still exist. This section reviews the existing literature on FER challenges, current trends in using social media data and synthetic data to enhance model generalizability, and identifies gaps that this paper seeks to address.

1.1 Challenges in facial emotion recognition

Facial emotion recognition (FER) has garnered extensive research interest due to its applications in human-computer interaction, psychological analysis, and marketing (Sariyanidi et al 2014). Traditional FER systems rely on large datasets of facial expressions to train models capable of recognizing a range of emotions (Rashed et al 2024). However, these datasets often lack diversity in demographics, cultural contexts, and emotional expressions, leading to models that perform inadequately when exposed to new content or different populations (Zhao et al 2011).

Datasets like CK+ (Lucey et al 2010) and JAFFE (Lyons et al 1998) are widely used but have limitations in terms of demographic diversity and expression variability (Sun et al 2023). Overfitting to specific datasets can result in biased models that do not capture the variability of facial expressions across different cultures, ages, and contexts (Cowen et al 2021). This limitation emphasizes the need for more diverse and representative datasets to improve the generalizability of FER models (Chutia and Baruah 2024).

1.2 Data augmentation techniques

To address the limitations of dataset diversity, data augmentation methods have been employed to artificially increase the size and variability of training data. Traditional augmentation techniques involve geometric transformations, such as rotation, scaling, and flipping, as well as photometric adjustments like brightness and contrast changes (Khalifa et al 2022). While these methods can improve model robustness to some extent, they often fail to introduce the level of diversity required for significant generalization improvements (Roth et al 2020).

Singh (2023) explored advanced augmentation strategies, including style transfer and adversarial examples, to enhance FER models. However, these approaches may still not capture the full range of real-world facial expressions and can introduce artifacts that negatively impact model performance.

1.3 Synthetic data generation

Synthetic data generation using techniques like Generative Adversarial Networks (GANs) (Goodfellow et al 2020) and Variational Autoencoders (VAEs) (Nissen et al 2021) has shown promise in enriching training datasets (Kim et al 2024). GANs, in particular, can generate realistic facial images with diverse expressions, poses, and lighting conditions (Alqahtani et al 2021).

Studies have demonstrated that incorporating synthetic data can improve FER model performance by providing samples of rare or underrepresented expressions (Qiu et al 2021). For instance, Antoniou et al (2018) introduced Data Augmentation GANs (DAGANs) to generate additional training data, resulting in improved classification accuracy. However, concerns remain regarding the realism of synthetic data and the potential for models to overfit to artifacts introduced during generation (Sixt et al 2018).

1.4 Leveraging social media data

Social media platforms like YouTube offer vast amounts of user-generated content that can serve as a rich source of diverse facial expressions and emotional reactions (Savchenko et al 2022). Incorporating social media data into FER training datasets can enhance diversity and improve model generalizability (Shin et al 2020).

Zhu et al (2021) created the Acted Facial Expressions in the Wild (AFEW) dataset by extracting frames from movies, capturing more natural expressions in unconstrained environments. Similarly, Ruensuk et al (2020) leveraged images from social media to construct a large-scale dataset for emotion recognition.

However, challenges such as data privacy, consent, annotation quality, and the unstructured nature of social media content need to be addressed (Di Minin et al 2021). The variability in video quality, lighting, and occlusions in social media data also poses difficulties for FER models (Li and Deng 2020).

1.5 Integrating real-world and synthetic data

Combining real-world data from social media with synthetic data generation techniques offers a promising approach to enhance dataset diversity and model generalizability. Some studies have begun to explore this integration.

Yan et al (2019) proposed a method that combines synthetic images generated by GANs with real images from existing datasets to train FER models. Their results indicated that the inclusion of synthetic data improved recognition rates. Similarly, Sukumar et al (2024) integrated synthetic data to address class imbalance in FER datasets, achieving better performance on minority classes.

Despite these efforts, there is still a gap in systematically integrating synthetic data with real-world social media data to address content generalizability comprehensively. Most studies focus on either synthetic data generation or the use of social media data independently. Further research is needed to understand how these data sources can be combined effectively to improve FER models.

1.6 Research gaps and objectives

While previous studies have addressed various aspects of dataset diversity and model generalizability in FER, there remains a need for approaches that integrate synthetic and real-world data to overcome data constraints comprehensively. This paper aims to fill this gap by demonstrating how synthetic data and social media data can be used to create more diverse and representative datasets for deep learning applications. In this study, we demonstrate FER in the context of interest/disinterest towards digital advertisements. This paper leverages novel datasets to compare FER models trained using data from a controlled experiment, data augmented with synthetic images generated using Generative AI, and real-world data extracted from YouTube. Through the inclusion of emotional responses to broader content and supplementing underrepresented areas in emotion categories with synthetic data, this research seeks to enhance the generalizability of FER models. The following research questions are addressed in this paper: (1) Can FER model generalizability be improved using more diverse data extracted from social media and/or generated using AI, as compared to controlled data from a laboratory study? (2) How can FER models trained on specific categories of advertisements be generalizable to new categories of advertisements?

2 Data description

In this study, we utilized three different data sources of human facial expressions: (1) data collected in the NeuroBioSense experiment (Kocaçınar et al 2024), (2) data we extracted from YouTube, and (3) data we created using a generative AI platform. This data was used to train three separate Xception-based Facial Expression Recognition (FER) models. Then, the three models were compared for performance using a subset of the NeuroBioSense data reserved for validation and subset of the NeuroBioSense reserved for testing.

These datasets were carefully designed to systematically assess the model's generalization capabilities across diverse participant demographics and various advertisement categories. The purpose of employing multiple datasets was to rigorously test the robustness and adaptability of the model in recognizing and interpreting facial expressions in different contexts. While these datasets are not equal in terms of size or proportion of interested/not-interested, our model metrics account for this imbalance and enable comparison across models.

2.1 NeuroBioSense (baseline) dataset

The first dataset, referred to as Baseline, is from the NeuroBioSense dataset (Kocaçınar et al 2024), and represents data collected in a controlled laboratory style environment. It consists of videos of facial reactions of participants while they were exposed to three distinct advertisement categories: (1) car and technology, (2) food and market, and (3) cosmetics and fashion. Each participant was assigned to one advertisement category, ensuring that the reactions were specific to that category. There were 20 participants assigned to car and technology, 20 to food and market, and 18 to cosmetics and fashion. Participants' facial expressions were captured in real-time and labeled based on their self-reported emotional responses, as interested or not-interested. This dataset includes a total of 58 participants

(30 female, 23 male), ages 18 to 66 (mean = 27.4, SD = 11.3). The structured nature of the dataset provided a consistent and controlled environment, allowing for a reliable baseline in the evaluation of the FER model.

The dataset includes a total of 1045 video recordings capturing participants' facial expressions while watching the various advertisements. These videos were selected for inclusion in our study in part because they are intentionally short, with an average duration of 2.78 s (range: 0.33–9.98 s). This short duration minimizes the likelihood of significant fluctuations in emotional expressions, thus making it reasonable to assign an overall emotional label ("Interested" or "Not Interested") to each video.

During the preprocessing phase, we systematically extracted frames from each of the 1045 video recordings to capture consistent facial expressions across the viewing of advertisements. Specifically, we extracted 10 frames from each video at equal intervals, determined by dividing the total number of frames by 10. This method ensured that frames were uniformly sampled throughout the entire duration of each video, which provides a representative temporal cross-section of facial expressions and avoids random selection to minimize sampling bias and enhancing the reliability of subsequent analyses. Thus, this resulted in a total of 10,450 images for our analysis. Specifically, there were 3020 images of participant faces while watching the car and technology ads (2,510 labeled as interested, 83%); 3370 images of participant faces while watching the food and market ads (1,860 labeled as interested, 55%); 4060 images of participant faces while watching the cosmetics and fashion ads (2,650 labeled as interested, 65%).

The baseline data was divided into a train set, validate set, and test set. The images of the participants watching the food/market were used for training, the images for cosmetics/fashion were used for validation, and the car/technology were saved for model testing. This was intentionally done to ensure that model validation and testing occurred on novel sets of people and ad content than the models were trained on.

The baseline train data was further augmented with the additional data sources, described below, for the other two models. However, all three models validated using the same 4060 images and tested using the same 3020 images.

2.2 NeuroBioSense and YouTube combined (Baseline + YouTube) dataset

To further diversify the dataset and enhance the model's ability to generalize, we expanded the NeuroBioSense data by incorporating a collection of YouTube reaction videos. These videos capture users' natural reactions to a variety of advertisements, providing a more dynamic and uncontrolled dataset compared to the original, which was collected in a more structured environment.

The YouTube reaction dataset focuses on participants' real-time emotional responses to advertisements, with reactions categorized as either interested or not-interested. This addition introduced several complexities absent in the NeuroBioSense dataset, including greater diversity in participants. Furthermore, the types of advertisements covered were broader, introducing a wider range of products and scenarios, which allowed the model to become more robust in handling real-world conditions.

To collect this data, we took screenshots of videos available on YouTube. We found these videos by searching for "reaction videos to advertisements," with the inclusion criteria that the video contained participants that were verbal, at some point, about their interest/disinter-

est in the ad content. In total, we used 137 different YouTube videos of 137 different people (44 female, 93 male) and tagged the videos manually as interested or not-interested based on what they said in the video. For each of the 137 people, there were on average 14 images (SD = 3.44 images) of them included in the dataset. As a result, the YouTube dataset adds an additional 2000 images to the baseline dataset. For these additional images, 1000 were tagged as interested and 1000 as not-interested.

This augmentation of the baseline data reflects an increase in diversity in both participant reactions and content exposure, helping to ensure that the model can generalize effectively across different user types and advertising contexts.

2.3 NeuroBioSense and generative AI-enhanced synthetic dataset (Baseline + StyleGAN2)

In the third dataset, we introduced synthetic data to further augment the baseline training data. This dataset was created using RunwayML, a generative AI platform that leverages advanced techniques like Style-based Generative Adversarial Networks (StyleGAN2) developed by NVIDIA, and designed for generating high quality realistic synthetic images of human faces (Karras et al 2020). The primary purpose of this synthetic data was to introduce even more diversity in facial features, such as hairstyle, race, background, and clothing, while maintaining consistency in emotional expressions.

The synthetic data was created using the labeled (interested or not-interested) NeuroBioSense training data. For each video in the original NeuroBioSense training data, we randomly extracted one frame for use in generating the synthetic data. For each of these frames, RunwayML generated six images, resulting in an additional 2022 images. This data was then incorporated into the training set to improve the model's ability to generalize across participants with unseen facial characteristics and features.

The use of generative AI allowed us to significantly increase the diversity of facial expressions and demographic variables without needing additional real-world data. This synthetic augmentation was instrumental in enhancing the model's ability to generalize across new faces, helping to overcome limitations encountered in conventional data augmentation techniques.

2.4 Dataset combinations for models

The three datasets described above were used to create the three conditions (Baseline, Baseline + YouTube, Baseline + StyleGAN2) for use in the FERs. Table 1 summarizes the number of images used in training, validating, and testing each model. Note, the validate and test sets were the same across all three models; only the training sets differed. Additionally, the test data consisted of a subset of participants and advertisement categories that were not seen in any training data.

The separate training and evaluation setups were designed to address the primary goal of this study which is to evaluate whether additional manual data (e.g., YouTube) or synthetic

Table 1 Number of images used in training, validating, and testing each model

Model	Train	Validate	Test	Total
Baseline	3370	4060	3020	10,450
Baseline + YouTube	5370	4060	3020	12,450
Baseline + StyleGAN2	5392	4060	3020	12,472

data (e.g., StyleGAN2) provides more value for improving model performance. By keeping the datasets separate, we could systematically assess the individual contributions of manual and synthetic data without conflating their effects. This modular approach allows us to derive actionable insights about the relative utility of different data sources. While combining all datasets for joint training could enhance data diversity and robustness, such an approach would make it difficult to isolate the specific impact of each dataset type. Evaluating manual and synthetic data independently ensures clarity in understanding their unique roles in augmenting the original dataset.

To ensure the robustness and generalizability of our models, we enhanced the diversity of our training datasets through the integration of social media and generative AI-generated data (Fig. 1). Specifically, our datasets comprise participants across a broad spectrum of demographics, including gender, racial/ethnic diversity, and age groups. This approach aligns with recommendations from prior research emphasizing the importance of demographic diversity for developing FER systems that can accurately interpret facial expressions across various populations (Cowen et al 2021). Integrating real-world data from social media platforms, such as YouTube, enables our models to capture context-rich emotional expressions that are often missing in controlled experimental setups (Shin et al 2020). Additionally, we used Generative Adversarial Networks (GANs) to supplement our dataset with synthetic images representing underrepresented groups, providing facial diversity that mirrors real-world conditions (Qiu et al 2021). Compared to traditional FER datasets like CK+ and JAFFE, which often lack sufficient demographic representation (Lucey et al 2010; Lyons et al 1998). Also, to enhance the contextual and cultural diversity of our dataset, we included a variety of clothing styles, such as traditional, casual, and formal attire from different cultures. Clothing serves as a visual cue, providing context for emotion interpretation by reflecting cultural identity, seasonality, and situational factors. This diversity reduces model bias and prevents overfitting to specific patterns present in limited datasets (Huang et al 2021).

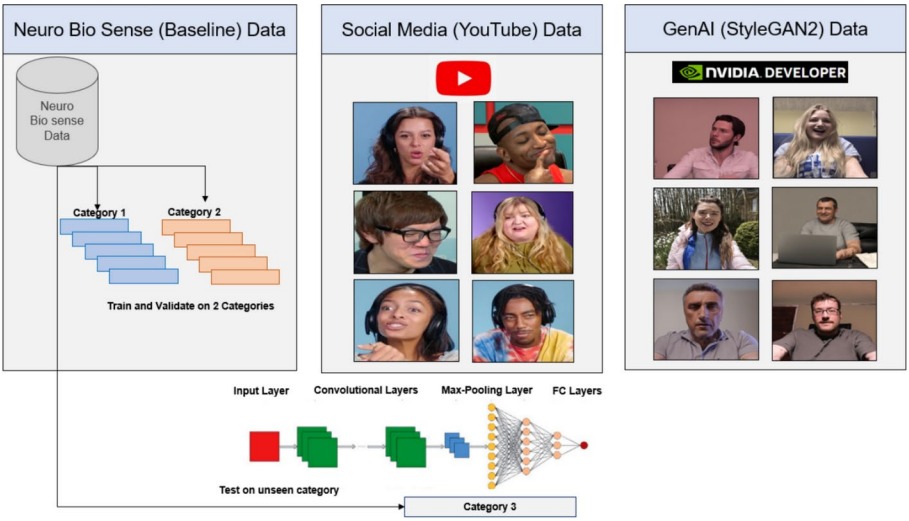

Fig. 1 Overview of the training and validating pipeline integrating NeuroBioSense (Baseline), Social Media (YouTube), and Generative AI (StyleGAN2) datasets

3 Theoretical framework: StyleGAN2

The resolution and quality of images produced by generative methods, especially generative adversarial networks (GAN), are improving rapidly (Li et al 2018). StyleGAN2 is a generative adversarial network (GAN) architecture that advances the capabilities of its predecessor, StyleGAN, by addressing specific limitations related to image quality and the presence of artifacts (Karras et al 2019). Developed by NVIDIA researchers (Karras et al 2020), StyleGAN2 introduces significant architectural modifications and new techniques to enhance the fidelity, realism, and controllability of synthesized images.

StyleGAN2 maintains specific facial details like expressions while altering other elements such as the background. This is achieved through its hierarchical architecture that modulates different aspects of the image at various layers (Melnik et al 2024). The model employs a mapping network that transforms an input latent vector into intermediate latent codes. These codes influence the generator's layers differently, with early layers controlling high-level attributes like facial structure and expressions, and later layers affecting fine-grained details like texture and background.

By keeping the latent codes consistent for early layers, StyleGAN2 preserves facial expressions. At the same time, it varies the codes for later layers, allowing background changes. This approach enables StyleGAN2 to disentangle features effectively (Bounareli et al 2024). As a result, it can keep details, like a consistent laugh on the face, while other aspects evolve independently. This is achieved through the model's style-based synthesis and progressive refinement mechanisms within the generator network (Chong et al 2021).

At the core of StyleGAN2 is the concept of style-based synthesis (Karmali et al 2022), where a mapping network transforms a latent vector $\mathrm{z} \in \mathcal{Z}$ into an intermediate latent space $\mathrm{w} \in \mathcal{W}$. This intermediate vector modulates the generator to produce images with disentangled attributes (Wu et al 2021).

In the original StyleGAN, the generator employed adaptive instance normalization (AdaIN) layers to inject style information (Huang and Belongie 2017). However, this approach led to characteristic artifacts, such as droplet-shaped distortions, due to inherent biases introduced by normalization operations (Bermano et al 2022). To eliminate these artifacts, StyleGAN2 replaces AdaIN with a novel mechanism called weight demodulation (Karras et al 2020).

Instead of normalizing the activations, StyleGAN2 modulates the weights of the convolutional layers directly based on the style vector. Mathematically, the modulation and demodulation processes are defined as follows:

The convolutional weights w_{ijk} are modulated by the style coefficients s_i:

$$\hat{w}_{ijk} = s_i \cdot w_{ijk}, \tag{1}$$

where i indexes the input feature maps, j the output feature maps, and k the spatial kernel positions.

To prevent the amplification of certain features and maintain consistent signal magnitudes, the modulated weights are demodulated:

$$\tilde{w}_{ijk} = \frac{\hat{w}_{ijk}}{\sqrt{\sum_i (s_i \cdot w_{ijk})^2 + \epsilon}}, \tag{2}$$

where ϵ is a small constant added for numerical stability.
The demodulation step effectively normalizes the weights, ensuring that the style modulation does not introduce undesired biases or artifacts (Yuan et al 2022). This approach maintains the statistical properties of the feature maps without relying on explicit normalization layers.

Another critical innovation in StyleGAN2 is the introduction of path length regularization (Karras et al 2020). This technique encourages the generator to produce images that respond smoothly and predictably to changes in the latent space $\mathcal{W}$. The regularization term R_{pl} is defined as:

$$R_{\text{pl}} = \mathbb{E}_{\text{w,y}\sim\mathcal{N}(0,I)} \left(\left\| \text{y}^\top \frac{\partial G(\text{w})}{\partial \text{w}} \right\|_2 - a \right)^2, \tag{3}$$

where $G(\text{w})$ is the generator output (image) given the latent vector w,y is a random vector sampled from a standard normal distribution and a is an exponential moving average of the path lengths to stabilize training.
This regularization penalizes deviations from the expected rate of change in the images with respect to the latent vectors, promoting a more linear and interpretable latent space.

StyleGAN2 also refines the mapping network and the hierarchical application of styles. By adjusting where and how the styles are applied within the generator, the architecture achieves a better separation of high-level attributes (such as pose and identity in face generation) from fine-grained details (like texture and color). The mapping network $f : \mathcal{Z} \to \mathcal{W}$ is designed to increase the expressiveness of the latent space $\mathcal{W}$, which enables more nuanced control over the generated images.

The discriminator in StyleGAN2 is augmented with techniques like lazy regularization (Karras et al 2020), where computationally intensive regularization terms are applied less frequently to reduce training overhead without sacrificing performance. The discriminator loss includes an R1 regularization term, which penalizes the gradient norm of the discriminator's output with respect to the input images:

$$R_1 = \frac{\gamma}{2} \mathbb{E}_{\text{x}\sim P_{\text{data}}} \left(\|\nabla_{\text{x}} D(\text{x})\|_2^2 \right), \tag{4}$$

where $D(\text{x})$ is the discriminator's output given real image x.γ is a regularization coefficient. P_{data} is the distribution of real images.
The culmination of these architectural and methodological advancements allows StyleGAN2 to generate images with unprecedented quality. The elimination of normalization biases, improved modulation techniques, and enhanced regularization contribute to the synthesis of images that are both highly realistic and controllable. The generator can produce high-resolution images (e.g., 1024×1024 pixels) that exhibit fine details (Fig. 2).

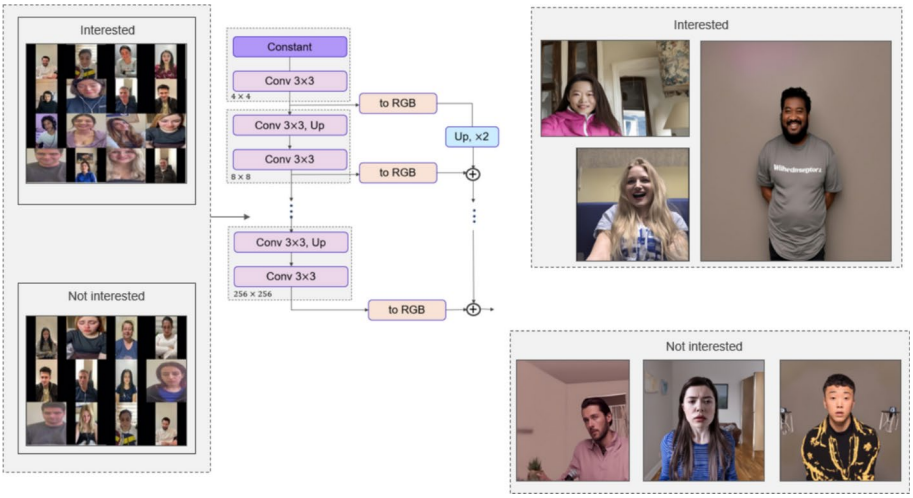

Fig. 2 Synthetic image generation using a StyleGAN model based on 'Interested' and 'Not Interested' image categories

4 Model training and optimization strategy

There were three separate models trained using the data combinations. The inclusion of diverse stimuli, such as advertisements across various categories, enhances the model's ability to generalize across different content types.

We formulated the FER task as a binary classification problem, distinguishing between expressions of interest and non-interest. The model optimization involves minimizing the binary cross-entropy loss function, defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{5}$$

where N is the number of samples, $y_i \in 0, 1$ is the true label, and p_i is the predicted probability that the ith sample belongs to the positive class.

We employed the Adam optimizer (Kingma and Ba 2015), which adapts the learning rate for each parameter using estimates of first and second moments of the gradients. The parameter updates are computed as:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{6}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{7}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{8}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{9}$$

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{10}$$

where g_t is the gradient at time step t, m_t and v_t are the exponentially decaying averages of past gradients and squared gradients, β_1 and β_2 are decay rates, α is the learning rate, and ϵ is a small constant to prevent division by zero.
The learning rate was initialized at $\alpha = 1 \times 10^{-4}$ and decayed exponentially to ensure convergence and prevent overfitting. Training was conducted over 300 epochs with a batch size of 16. After finalizing the optimization process and model parameters, we utilized Google Cloud's e2-highmem-16 instance equipped with an NVIDIA Tesla P100 GPU to handle the computational demands of training the Convolutional Neural Networks.

4.1 Data augmentation and regularization

To enhance model robustness and generalization, we applied various data augmentation techniques to the training data (Lewy and Mańdziuk 2023). First, we implemented geometric transformations, including random rotations with angles θ in the range $[-15°, 15°]$, horizontal flips, and random cropping to simulate different viewing angles and facial orientations. Additionally, photometric adjustments such as random changes in brightness, contrast, and saturation were applied to mimic varying lighting conditions. To account for sensor noise, Gaussian noise with zero mean and small variance was added to the images (Lewy and Mańdziuk 2023).

Regularization techniques were also employed to prevent overfitting. We applied dropout regularization with a dropout rate of 0.5 to the fully connected layers, randomly deactivating neurons during training (Srivastava et al 2014). Furthermore, weight decay, also known as L2 regularization, was incorporated by adding a penalty term $\lambda \|\theta\|_2^2$ to the loss function, where λ is the regularization parameter and θ represents the model weights (Moradi et al 2020). This penalty discourages the network from assigning excessively large weights, promoting simpler models that generalize better.

5 Experimental design

The experimental design of this study aims to assess the generalizability of FER models by using data from diverse sources, including both real-world and synthetic datasets. To ensure a robust and fair evaluation of model performance, a category-based data splitting strategy was adopted. This approach ensures that the training, validation, and test subsets are separated based on advertisement categories, allowing the models to be trained and validated on specific categories while being tested on entirely new category.

The rationale for using category-based data splitting lies in the need to evaluate the model's ability to generalize beyond the categories it has been trained on. This approach mimics real-world scenarios where FER systems encounter entirely new content types. By training

on one set of categories and testing on different, unseen categories, the design measures the model's capacity to adapt to novel contexts, a critical aspect of enhancing model robustness.

The category-based splitting approach offers several advantages for evaluating the FER models. First, it provides a realistic assessment of the model's ability to handle new advertisement categories, reflecting real-world deployment scenarios where models often encounter unfamiliar stimuli. Second, this strategy reduces potential biases that could arise from overlapping content between training, validation, and test sets, ensuring that the model's performance reflects true generalizability rather than overfitting to specific patterns.

6 FER model framework

6.1 Xception

In this study, we employ the Xception model, a convolutional neural network (CNN) architecture that has demonstrated superior performance in previous research by Alipour et al (2024) using the same NeuroBioSense data. The Xception model builds upon the Inception architecture (Poma et al 2020) by replacing the standard Inception modules with depthwise separable convolutions, thereby enhancing both computational efficiency and model accuracy. Our selection of the Xception architecture is motivated by its ability to optimize resource utilization while achieving high performance in large-scale visual recognition tasks (Alzubaidi et al 2024), making it particularly suitable for FER in diverse content environments.

6.2 Xception model architecture

The term Xception stands for Extreme Inception (Chollet 2017), reflecting its evolution from the original Inception model. The key innovation in Xception is the use of depthwise separable convolutions, which decompose a standard convolution into a depthwise convolution and a pointwise convolution. This decomposition allows for a more efficient representation of the convolutional operation, reducing the number of parameters and computational cost without sacrificing model capacity.

A standard convolutional operation combines spatial filtering and channel mixing in a single step. Mathematically, a standard convolution for an input tensor $\mathrm{X} \in \mathbb{R}^{h \times w \times c_{\mathrm{in}}}$ with c_{in} input channels, applying c_{out} filters of size $k \times k$, produces an output tensor $\mathrm{Y} \in \mathbb{R}^{h' \times w' \times c_{\mathrm{out}}}$, computed as:

$$\mathrm{Y}i, j, k = \sum u = 1^{k} \sum_{v=1}^{k} \sum_{c=1}^{c_{\mathrm{in}}} \mathrm{W}u, v, c, k \cdot \mathrm{X}i + u - 1, j + v - 1, c \tag{11}$$

where $\mathrm{W} \in \mathbb{R}^{k \times k \times c_{\mathrm{in}} \times c_{\mathrm{out}}}$ represents the convolutional filters.
In contrast, a depthwise separable convolution splits this operation into two separate layers of depthwise and pointwise.

Depthwise convolution which applies a single convolutional filter per input channel independently.

$$\mathrm{Z}_{i,j,c} = \sum_{u=1}^{k} \sum_{v=1}^{k} \mathrm{K}_{u,v,c} \cdot \mathrm{X}_{i+u-1,j+v-1,c} \tag{12}$$

where $\mathrm{K} \in \mathbb{R}^{k \times k \times c_{\mathrm{in}}}$ is the depthwise convolutional filter.
Also, the pointwise convolution which applies a 1×1 convolution to combine the outputs of the depthwise convolution across channels.

$$\mathrm{Y}_{i,j,k} = \sum_{c=1}^{c_{\mathrm{in}}} \mathrm{P}_{c,k} \cdot \mathrm{Z}_{i,j,c} \tag{13}$$

where $\mathrm{P} \in \mathbb{R}^{c_{\mathrm{in}} \times c_{\mathrm{out}}}$ represents the pointwise convolutional filters.
This separation reduces the computational complexity significantly lowering the number of parameters and operations.

The Xception architecture is organized into three main parts of entry flow, middle flow, and exit flow. The entry flow is responsible for capturing low-level features and reducing the spatial dimensions of the input data. It typically uses convolutional layers followed by max-pooling operations to down-sample the data.

Next, the middle flow is made up of several identical modules (repeated eight times) designed to learn increasingly complex features. These modules contain depthwise separable convolution layers, which are more efficient than standard convolutions, and they also use residual connections to help preserve information as it passes through the layers.

Finally, the exit flow focuses on extracting high-level features and preparing the data for classification. This phase also uses depthwise separable convolutions, followed by global average pooling to reduce the dimensions further, and a fully connected layer for producing the final output suitable for classification tasks.

Residual connections are incorporated to mitigate the vanishing gradient problem and facilitate the training of deeper networks (Liu et al 2022). The overall architecture enables the model to learn rich feature representations essential for distinguishing subtle facial expressions in FER tasks.

6.3 Implementation details

The model was implemented using the Keras library with a TensorFlow backend. The top layers were replaced with a global average pooling layer followed by a fully connected layer with a sigmoid activation function for binary classification and early stopping was implemented based on the validation loss to prevent overfitting.

6.4 Metrics for evaluating model performance

To assess the effectiveness of our models in recognizing facial expressions and determining interest or non-interest, we employ two primary metrics: loss and accuracy (Goceri 2024). These metrics offer detailed understanding of the models' learning progress, capacity to generalize to novel data, and overall robustness (Liu et al 2024). The combination of these

metrics provides a balanced view of model performance across diverse datasets, highlighting how well the models have learned and adapted effectively to varying data patterns.

6.4.1 Loss function

The loss function provides a scalar value that represents the cost associated with the network's predictions (Alzahrani et al 2024). During training, the objective is to find the set of network parameters (weights and biases) that minimize this loss. This process is carried out using optimization algorithms; Adaptive Moment Estimation and stochastic gradient descent and its variants, which rely on the gradient of the loss function with respect to the network parameters (Krizhevsky et al 2012).

6.4.2 Accuracy

Accuracy quantifies the proportion of correct predictions made by the model out of all predictions which provides a straightforward measure of how well the model generalizes to unseen data. Mathematically, for a dataset with n samples, accuracy A is calculated as:

$$A = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i = \hat{y}_i)$$

where y_i is the true label of the ith sample. $\hat{y}_i$ is the predicted label of the ith sample. $\mathbb{I}$ is the indicator function, which returns 1 if the argument is true and 0 otherwise.
Accuracy serves as a primary metric during both training and evaluation phases of convolutional neural network models. It provides an immediate sense of the model's ability to correctly classify input data. High accuracy indicates effective learning of the underlying patterns in the data, while low accuracy suggests the need for model refinement.

7 Results and discussion

This section presents a comparative analysis of three models trained on distinct datasets: Baseline, Baseline + YouTube, and Baseline + StyleGAN2. The models are evaluated based on key performance metrics, including loss, accuracy, precision–recall curves, and ROC curves, providing discernment into their robustness, generalizability, and adaptability to diverse data sets.

7.1 Loss

Figure 3 illustrates the progression of training and validation losses across 300 epochs for each of the models to provide information about their learning dynamics and generalization capabilities. The Baseline + YouTube model (green line) demonstrates the most consistent validation loss, indicative of its capacity to generalize effectively to unseen data. Its training loss also remains low and aligns closely with the validation loss which shows that the model effectively learns from the diverse, real-world data without overfitting. The Baseline

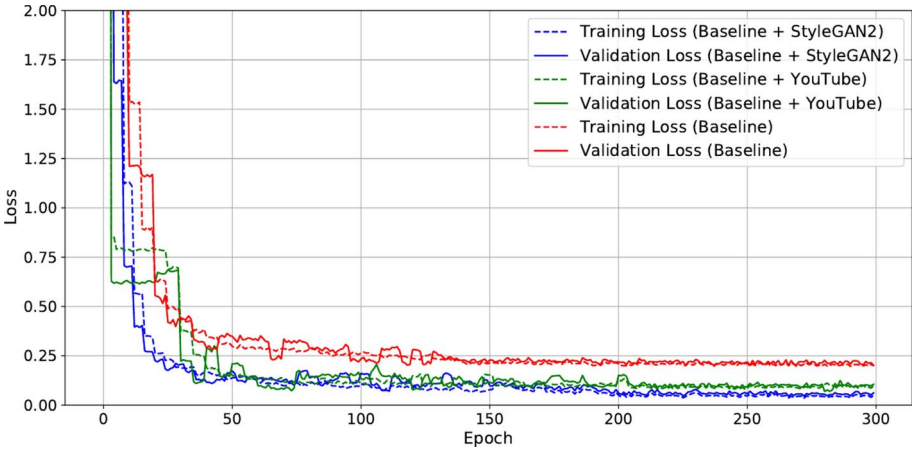

Fig. 3 Comparison of training and validation losses across the three datasets over 300 epochs. (Color figure online)

+ StyleGAN2 model (blue line) initially displays a higher loss which suggests challenges in adapting to synthetic data, but ultimately converges. It highlights the potential of synthetic augmentation to effectively complement real-world data and eventually enhance model performance. Conversely, the Baseline model (red line) exhibits rapid initial reduction in training loss during initial epochs that reflects its capacity to quickly adapt to the limited dataset, yet suffers from elevated validation loss that signifies limited generalization due to the constraints of its dataset. The persistent gap between its training and validation losses suggests overfitting, resulting from the constrained diversity of the training data, which limits the model's capacity to handle varied, real-world expressions.

7.2 Accuracy

During training, accuracy is also monitored on both training and validation datasets to assess the model's learning progress and generalization ability. Figure 4 illustrates the trends in training and validation accuracy over 300 epochs for each model and it provides insights into their learning efficiency and generalization capabilities. The training accuracy indicates how well the models learn from their respective datasets, while validation accuracy measures how effectively the models generalize to unseen data.

The Baseline + YouTube model, represented in green, demonstrates a steady increase in training accuracy over the epochs that shows effective learning from the diverse, real-world data. Notably, the training accuracy aligns closely with validation accuracy which shows the model avoids overfitting and maintains robust generalization capabilities. This alignment can be attributed to the inclusion of varied expressions and contexts from the YouTube dataset, which exposes the model to a wide range of real-world patterns during training and enables it to handle unseen data effectively.

In contrast, the Baseline + StyleGAN2 model, depicted in blue, achieves rapid improvements in training accuracy it reaches high levels early in the training process. This indicates that the model quickly adapts to the synthetic patterns present in the data. While this suggests effective utilization of synthetic augmentation, it also raises concerns about potential

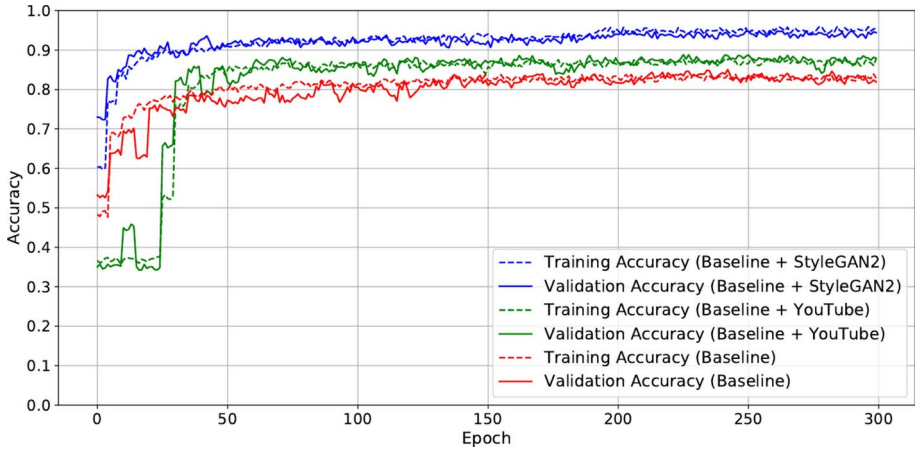

Fig. 4 Comparison of training and validation accuracies across three datasets over 300 epochs. (Color figure online)

overfitting, as evidenced by a noticeable gap between training and validation accuracy. The validation accuracy, although improving over time, remains lower than training accuracy, suggesting that the synthetic data introduces patterns that are less representative of real-world variability. Consequently, the model's ability to generalize effectively is limited by the artificial structure inherent to synthetic data, which may not capture the complexity of natural expressions as accurately as real-world data does.

The Baseline model, shown in red, displays rapid gains in training accuracy, reflecting its quick adaptation to the limited patterns of the constrained dataset. However, this strong performance during training does not translate well to the validation set, where the accuracy remains significantly lower. The persistent gap between training and validation accuracy highlights the model's limited capacity for generalization, which is likely due to the homogeneity and restricted diversity of the dataset. As a result, the model tends to overfit to the specific patterns present in the training data, failing to adapt to varied, unseen expressions.

7.3 Precision–recall and receiver operating characteristic curve

Beyond the evaluation of loss and accuracy, the models' performance is further assessed using the precision–recall (PR) curve and the receiver operating characteristic (ROC) curve. These metrics offer a more nuanced and detailed perspective, particularly in contexts where class imbalance may render accuracy alone an insufficient measure of model performance.

The PR curve (Fig. 5—left) measures the trade-off between precision and recall across various decision thresholds. Precision, or positive predictive value, reflects the proportion of correctly identified positive cases among all predicted positives, while recall, or true positive rate, denotes the proportion of correctly identified positive cases among all actual positives. This metric is particularly informative in contexts where class imbalances are present, as it helps highlight the model's capability to minimize false positives while maintaining high recall. As depicted in Fig. 5—left, the Baseline + StyleGAN2 model achieves the highest area under the PR curve (AUC = 0.94), indicating superior balance between precision and recall. The Baseline + YouTube model follows closely with an AUC of 0.89, suggesting

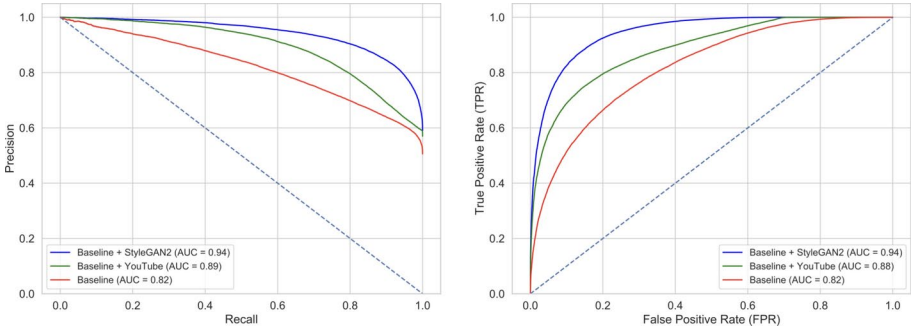

Fig. 5 PR (left) and ROC (right) curves illustrating models' performance and balance across different datasets

Table 2 Performance improvements for PR AUC and ROC-AUC across models/datasets

Model/Dataset	PR AUC	ROC AUC	Improvement from Baseline
Baseline	0.82	0.82	–
Baseline + YouTube	0.89	0.88	+ 0.07 (PR), + 0.06 (ROC)
Baseline + StyleGAN2	0.94	0.94	+ 0.12 (PR), + 0.12 (ROC)

strong generalizability to real-world data. In contrast, the Baseline model, with an AUC of 0.82, demonstrates weaker performance, suggesting challenges in maintaining both precision and recall due to limited data diversity.

The ROC curve (Fig. 5—right) provides another perspective by plotting the true positive rate against the false positive rate across varying thresholds. The Area Under the Curve (AUC) serves as a summary metric, reflecting the model's overall ability to distinguish between positive and negative classes. In Fig. 5—right, the Baseline + StyleGAN2 model again achieves the highest ROC-AUC at 0.94 which indicates a strong discriminative capacity. The Baseline + YouTube model, with an AUC of 0.88, also exhibits robust performance which confirms its effectiveness in generalizing to diverse real-world data. The Baseline model, however, achieves a lower AUC of 0.82, highlighting its limited ability to effectively separate classes, which can be attributed to the constraints of its dataset.

In particular, the PR and ROC curves were chosen as they comprehensively represent model performance across all classification thresholds and offer a detailed visualization of trade-offs. PR AUC is particularly valuable in scenarios with imbalanced datasets as it emphasizes the performance of the minority class to ensure that the model's predictions are reliable even when positive instances are scarce. Conversely, ROC AUC provides a broader perspective by balancing sensitivity and specificity and making it a complementary metric to PR AUC. This dual approach ensures a robust evaluation of model performance.

Additionally, Table 2 highlights significant percentage improvements in the PR AUC and ROC-AUC metrics following the integration of diverse data sources into the FER models. Initially, the Baseline model achieved a PR AUC of 0.82. With the inclusion of real-world YouTube data (Baseline + YouTube), this metric rose to 0.89, representing an approximate 8.5% improvement over the baseline. This gain indicates a more balanced model performance, with better detection of true positives and reduced false positives. When synthetic data from StyleGAN2 (Baseline + StyleGAN2) was added, the PR AUC increased further to

0.94, reflecting a 14.6% improvement over the baseline. This notable enhancement suggests that the synthetic data effectively addressed the variability in facial expressions, filling gaps left by the real-world dataset. Similarly, the ROC-AUC for the Baseline model was initially 0.82, which increased to 0.88 with YouTube data, showing a 7.3% improvement. This indicates stronger discriminative power, as the model became more adept at distinguishing between positive and negative classes across a broader spectrum of emotional expressions. With the integration of synthetic data in the Baseline + StyleGAN2 model, the ROC-AUC reached 0.94, marking a 14.6% improvement.

Lastly, to further understand the performance differences among the three models, we conducted an Analysis of Variance (ANOVA) on the AUC scores obtained from both the ROC and PR curves. The ANOVA test aims to assess whether the mean differences in AUC scores across the models are statistically significant.

The ANOVA results revealed a statistically significant difference in the mean AUC scores across the three models ($F(2, 297) = 24.38,\ p < 0.001$). This indicates that the inclusion of YouTube and synthetic data led to significant improvements in model generalizability, as reflected in the AUC scores.

8 Ethical considerations and solutions

The integration of both real-world and synthetic data in FER models holds significant potential for enhancing model performance and generalizability. However, it also raises complex ethical issues related to bias, fairness, and privacy (Bhanot et al 2021). This section examines these ethical challenges, discussing potential biases in FER models, privacy concerns arising from social media data usage (Anzum et al 2022), and proposing solutions to ensure the responsible development and deployment of FER systems. By addressing these ethical considerations, the study aims to contribute to the creation of more inclusive and fair FER models (Li et al 2021).

8.1 Biases in facial emotion recognition (FER) models

FER models have gained significant traction across domains such as marketing, mental health assessment, and human-computer interaction (Ekundayo and Viriri 2021). Despite their potential, these models are susceptible to biases arising from the limited diversity of training datasets, which can result in skewed interpretations of facial expressions (Sajjad et al 2023). A common issue is the over representation or under representation of specific demographic groups, such as certain races, age groups, or genders, leading to model inaccuracies when applied to broader populations (Dominguez-Catena et al 2024). While this study aims to address these biases by integrating synthetic data, it is critical to observe that synthetic data generation itself may replicate and even amplify biases inherent in the original datasets (Mannino and Abouzied 2019). For instance, GANs, which are used for creating synthetic data, can inadvertently perpetuate imbalances (Schneider 2024), especially if the initial data contains disproportionate samples from specific groups (Lilienthal 2024). This tendency may affect the fairness and accuracy of FER models when deployed in real-world applications, making it imperative to examine and mitigate potential biases in both real-world and synthetic data sources (Humphreys et al 2024).

8.2 Privacy concerns in social media data

The use of social media data to enhance FER models introduces complex ethical challenges, particularly concerning privacy, consent, and data protection (Mattioli and Cabitza 2024). Social media platforms offer a vast pool of user-generated content that can provide valuable diversity in facial expressions. However, this diversity comes with ethical dilemmas related to the rights of users whose content is used without explicit consent (Stommel and Rijk 2021). While social media data presents opportunities for improving model generalizability, it also poses risks of violating user privacy, as individuals may be unaware of their content being utilized for research or commercial purposes (Van der Schyff et al 2020). Additionally, the unstructured nature of social media data, which often includes personal information, raises ethical concerns about the extent to which data can be anonymized without compromising its utility for training models (Rossi et al 2022). Therefore, it is crucial to reflect on the ethical implications of data use, particularly in balancing the benefits of model enhancement with the rights of individuals to privacy and data protection.

8.3 Privacy concerns in synthetic data

While synthetic data generation offers a promising solution for augmenting datasets (Paulin and Ivasic-Kos 2023) and alleviating privacy concerns, it also presents unique ethical and privacy challenges (Hewage et al 2023). Synthetic data, particularly from Generative Adversarial Networks, can inadvertently reproduce identifiable patterns from original datasets (Ma et al 2024), risking privacy and potentially perpetuating inherent biases. This issue is especially prominent with smaller datasets, where GANs may overfit on particular features, resulting in synthetic data that resembles real individuals with poor quality (Rather et al 2024). By addressing these privacy challenges, researchers can leverage synthetic data's benefits.

8.4 Addressing ethical issues and promoting fairness

To foster ethical integrity in FER model development, it is essential to establish robust measures that address bias mitigation and privacy protection (Butt et al 2023). First, conducting comprehensive data audits can play a vital role in identifying and correcting imbalances in demographic representation (Kim et al 2021). By analyzing demographic distributions in both real-world and synthetic datasets, researchers can ensure that training data aligns more closely with the intended deployment population (Humphreys et al 2024). This process should be coupled with the adoption of fairness-aware generative models, such as FairGAN (Li et al 2022) or DebiasGAN (Sinha et al 2020), which aim to produce synthetic data that better represents underrepresented groups without exaggerating biases found in the original data.

In addition to addressing biases in training data, it is equally important to consider the ethical implications of using social media and synthetica AI generated data (Schmitt and Flechais 2024). One solution is to ensure transparency in data collection processes. This involves using datasets that are licensed for research, prioritizing those where contributors have provided explicit consent for their data to be used (Lunnay et al 2015). Where consent is not feasible, researchers should employ anonymization techniques that protect individ-

ual identities while preserving the diversity of expressions needed for FER model training (Pawar et al 2018). Establishing clear protocols for obtaining permission and using ethically sourced datasets is crucial in maintaining the integrity of FER research.

Furthermore, integrating fairness metrics during model training can enhance the equity of FER models. Metrics such as Equalized Odds (Romano et al 2020) and Demographic Parity (Jiang et al 2022) can help evaluate whether the model performs consistently across demographic groups, thus promoting fairness and minimizing the risk of discriminatory outcomes. It is imperative to implement these metrics throughout the model development process, from training to validation, ensuring that the model's generalization capabilities extend fairly across different demographic contexts.

Lastly, ethical compliance should extend beyond technical adjustments. Adherence to global data protection regulations, such as the General Data Protection Regulation (GDPR) (Voigt and Von dem Bussche 2017) and the California Consumer Privacy Act (CCPA) (Harding et al 2019), must guide all aspects of data handling. Researchers should collaborate with legal experts and ethics review boards to ensure compliance with these regulations, thereby aligning FER model development with legal standards of privacy protection (Dabis and Csáki 2024)

8.5 Future ethical directions in FER research

Looking ahead, it is essential to adopt more inclusive data collection strategies that are rooted in explicit consent and ethical sourcing. The implementation of differential privacy techniques holds promise for safeguarding individual privacy while allowing models to learn from data effectively. Moreover, advancing explainable AI (XAI) techniques in FER could enhance transparency which enables users to understand the basis of emotion predictions and identify potential biases in real-time (Speith 2022). Future research should prioritize these approaches, aiming to establish FER models that are both technologically robust and ethically sound.

9 Conclusion

In this study, we set out to improve the generalizability of an FER convolutional neural network Xception model by integrating diverse data sources. We combined real-world data from YouTube, controlled experimental data, and synthetic images generated by StyleGAN2.

The results demonstrate that integrating these diverse datasets significantly improves model performance. The models trained with the combined data sources consistently outperformed those trained on individual datasets, showing better accuracy and reduced overfitting. The addition of real-world data from YouTube enriched the models with more natural variations of facial expressions, closely reflecting real-world dynamics. Meanwhile, the synthetic data generated by StyleGAN2 filled critical gaps, introducing underrepresented facial expressions and enhancing the diversity of the training dataset.

These findings highlight the importance of data diversity in achieving model generalizability. However, this process also highlighted several ethical considerations. The use of social media data raises privacy concerns, particularly regarding consent and data protection, while synthetic data carries the risk of reinforcing existing biases if not properly

managed. Thus, the study not only contributes to improving FER models technically but also emphasizes the need for responsible AI practices that balance innovation with ethical integrity.

In conclusion, while the integration of real-world and synthetic data represents a promising strategy for enhancing FER models, it also necessitates continuous refinement to ensure both technical performance and ethical compliance. Future research should focus on further improving the realism of synthetic data and exploring ways to ensure fairness and transparency in FER systems. This work thus lays the groundwork for developing FER models that are not only technically advanced but also ethically robust and capable of performing reliably across diverse, real-world scenarios.

Acknowledgements The authors would like to acknowledge Google Cloud Research Team for providing the computational resources used in this study.

Author contributions P.A. performed the analysis and prepared all figures/tables. E.G. provided supervision on the project. All authors wrote and edited the main manuscript text.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors report there are no conflict of interest to declare.

References

Algabri R, Abdu A, Lee S (2024) Deep learning and machine learning techniques for head pose estimation: a survey. Artif Intell Rev 57(10):1–66

Alipour P, Gallegos EE, Sridhar S (2024) AI-driven marketing personalization: deploying convolutional neural networks to decode consumer behavior. Int J Hum Comput Interact. https://doi.org/10.1080/10447318.2024.2432455

Alqahtani H, Kavakli-Thorne M, Kumar G (2021) Applications of generative adversarial networks (GANs): an updated review. Arch Comput Methods Eng 28:525–552

Alzahrani M, Usman M, Jarraya SK et al (2024) Deep models for multi-view 3d object recognition: a review. Artif Intell Rev 57(12):1–71

Alzubaidi L, Fadhel MA, Hollman F et al (2024) SSP: self-supervised pertaining technique for classification of shoulder implants in X-ray medical images: a broad experimental study. Artif Intell Rev 57(9):261

Antoniou A, Storkey A, Edwards H (2018) Augmenting image classifiers using data augmentation generative adversarial networks. In: Artificial neural networks and machine learning—ICANN 2018: 27th International conference on artificial neural networks, Rhodes, Greece, 4–7 October 2018, Proceedings, Part III 27. Springer, pp 594–603

Anzum F, Asha AZ, Gavrilova ML (2022) Biases, fairness, and implications of using AI in social media data mining. In: 2022 International conference on cyberworlds (CW). IEEE, pp 251–254
Bermano AH, Gal R, Alaluf Y et al (2022) State-of-the-art in the architecture, methods and applications of StyleGAN. In: Computer graphics forum. Wiley Online Library, pp 591–611
Bhanot K, Qi M, Erickson JS et al (2021) The problem of fairness in synthetic healthcare data. Entropy 23(9):1165
Bounareli S, Tzelepis C, Argyriou V et al (2024) One-shot neural face reenactment via finding directions in GAN's latent space. Int J Comput Vis 132:3324–3335
Butt MA, Qayyum A, Ali H et al (2023) Towards secure private and trustworthy human-centric embedded machine learning: an emotion-aware facial recognition case study. Comput Secur 125:103058
Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
Chong MJ, Chu WS, Kumar A et al (2021) Retrieve in style: Unsupervised facial feature transfer and retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3887–3896
Chutia T, Baruah N (2024) A review on emotion detection by using deep learning techniques. Artif Intell Rev 57(8):203
Cowen AS, Keltner D, Schroff F et al (2021) Sixteen facial expressions occur in similar contexts worldwide. Nature 589(7841):251–257
Dabis A, Csáki C (2024) AI and ethics: investigating the first policy responses of higher education institutions to the challenge of generative AI. Hum Soc Sci Commun 11(1):1–13
Di Minin E, Fink C, Hausmann A et al (2021) How to address data privacy concerns when using social media data in conservation science. Conserv Biol 35(2):437–446
Dominguez-Catena I, Paternain D, Jurio A et al (2024) Less can be more: representational vs. stereotypical gender bias in facial expression recognition. Prog Artif Intell. https://doi.org/10.1007/s13748-024-00345-w
Ekundayo OS, Viriri S (2021) Facial expression recognition: a review of trends and techniques. IEEE Access 9:136944–136973
Goceri E (2024) Gan based augmentation using a hybrid loss function for dermoscopy images. Artif Intell Rev 57(9):234
Goodfellow I, Pouget-Abadie J, Mirza M et al (2020) Generative adversarial networks. Commun ACM 63(11):139–144
Harding EL, Vanto JJ, Clark R et al (2019) Understanding the scope and impact of the California consumer privacy act of 2018. J Data Prot Privacy 2(3):234–253
Hewage U, Sinha R, Naeem MA (2023) Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review. Artif Intell Rev 56(9):10427–10464
Huang X, Belongie S (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision, pp 1501–1510
Huang Y, Wu Q, Xu J et al (2021) Clothing status awareness for long-term person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11895–11904
Humphreys D, Koay A, Desmond D et al (2024) AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business. AI Ethics 4:791–804
Jiang Z, Han X, Fan C et al (2022) Generalized demographic parity for group fairness. In: International conference on learning representations
Karmali T, Parihar R, Agrawal S et al (2022) Hierarchical semantic regularization of latent spaces in StyleGANs. In: European conference on computer vision. Springer, pp 443–459
Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4401–4410
Karras T, Laine S, Aittala M et al (2020) Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8110–8119
Khalifa NE, Loey M, Mirjalili S (2022) A comprehensive survey of recent trends in deep learning for digital images augmentation. Artif Intell Rev 55(3):2351–2377
Kim E, Bryant D, Srikanth D, et al (2021) Age bias in emotion detection: an analysis of facial emotion recognition performance on young, middle-aged, and older adults. In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society, pp 638–644
Kim JJ, Srivatsa AV, Nahass GR et al (2024) Generative AI can effectively manipulate data. AI Ethics. https://doi.org/10.1007/s43681-024-00630-3
Kingma DP, Ba J (2015) ADAM: a method for stochastic optimization 3rd international conference on learning representations. In: ICLR 2015—conference track proceedings
Kocaçınar B, İnan P, Zamur EN et al (2024) Neurobiosense: a multidimensional dataset for neuromarketing analysis. Data Brief 53:110235
Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, vol 25. ACM

Lewy D, Mańdziuk J (2023) An overview of mixing augmentation methods and augmentation strategies. Artif Intell Rev 56(3):2111–2169
Li S, Deng W (2020) Deep facial expression recognition: a survey. IEEE Trans Affect Comput 13(3):1195–1215
Li H, Chen H, Li B et al (2018) Can forensic detectors identify GAN generated images? In: 2018 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC). IEEE, pp 722–727
Li Y, Gao Y, Chen B et al (2021) Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild. IEEE Trans Circuits Syst Video Technol 32(5):3190–3202
Li J, Ren Y, Deng K (2022) FairGAN: GANs-based fairness-aware learning for recommendations with implicit feedback. In: Proceedings of the ACM web conference, vol 2022, pp 297–307
Lilienthal DB (2024) Synthetic data generation for accurate, fair, and private recommender systems. Master's thesis, San Jose State University
Liu Y, Zhang L, Hao Z et al (2022) An xception model based on residual attention mechanism for the classification of benign and malignant gastric ulcers. Sci Rep 12(1):15365
Liu P, Qian W, Zhang H et al (2024) Automatic sleep stage classification using deep learning: signals, data representation, and neural networks. Artif Intell Rev 57(11):301
Lucey P, Cohn JF, Kanade T et al (2010) The extended Cohn–Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE, pp 94–101
Lunnay B, Borlagdan J, McNaughton D et al (2015) Ethical use of social media to facilitate qualitative research. Qual Health Res 25(1):99–109
Lyons M, Akamatsu S, Kamachi M et al (1998) Coding facial expressions with GABOR wavelets. In: Proceedings of 3rd IEEE international conference on automatic face and gesture recognition. IEEE, pp 200–205
Ma Z, Mei G, Xu N (2024) Generative deep learning for data generation in natural hazard analysis: motivations, advances, challenges, and opportunities. Artif Intell Rev 57(6):160
Mannino M, Abouzied A (2019) Is this real? generating synthetic data that looks real. In: Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology, pp 549–561
Mattioli M, Cabitza F (2024) Not in my face: Challenges and ethical considerations in automatic face emotion recognition technology. Machine Learning and Knowledge Extraction 6(4):2201–2231
Melnik A, Miasayedzenkau M, Makaravets D et al (2024) Face generation and editing with StyleGAN: a survey. IEEE Trans Pattern Anal Mach Intell 46(5):3557–3576
Moradi R, Berangi R, Minaei B (2020) A survey of regularization strategies for deep models. Artif Intell Rev 53(6):3947–3986
Nissen JN, Johansen J, Allesøe RL et al (2021) Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol 39(5):555–560
Paulin G, Ivasic-Kos M (2023) Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. Artif Intell Rev 56(9):9221–9265
Pawar A, Ahirrao S, Churi PP (2018) Anonymization techniques for protecting privacy: a survey. In: 2018 IEEE Punecon. IEEE, pp 1–6
Poma XS, Riba E, Sappa A (2020) Dense extreme inception network: towards a robust CNN model for edge detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1923–1932
Qiu H, Yu B, Gong D et al (2021) Synface: face recognition with synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10880–10890
Rashed AEE, Atwa AEM, Ahmed A et al (2024) Facial image analysis for automated suicide risk detection with deep neural networks. Artif Intell Rev 57(10):1–43
Rather IH, Kumar S, Gandomi AH (2024) Breaking the data barrier: a review of deep learning techniques for democratizing AI with small datasets. Artif Intell Rev 57(9):226
Romano Y, Bates S, Candes E (2020) Achieving equalized odds by resampling sensitive attributes. Adv Neural Inf Process Syst 33:361–371
Rossi A, Arenas MP, Kocyigit E et al (2022) Challenges of protecting confidentiality in social media data and their ethical import. In: 2022 IEEE European symposium on security and privacy workshops (EuroS &PW). IEEE, pp 554–561
Roth K, Milbich T, Sinha S et al (2020) Revisiting training strategies and generalization performance in deep metric learning. In: International conference on machine learning, PMLR, pp 8242–8252
Ruensuk M, Cheon E, Hong H et al (2020) How do you feel online: exploiting smartphone sensors to detect transitory emotions during social media use. In: Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, vol 4(4), pp 1–32
Sajjad M, Ullah FUM, Ullah M et al (2023) A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. Alex Eng J 68:817–840

Sariyanidi E, Gunes H, Cavallaro A (2014) Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans Pattern Anal Mach Intell 37(6):1113–1133
Savchenko AV, Savchenko LV, Makarov I (2022) Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Trans Affect Comput 13(4):2132–2143
Schmitt M, Flechais I (2024) Digital deception: generative artificial intelligence in social engineering and phishing. Artif Intell Rev 57(12):1–23
Schneider J (2024) Explainable generative AI (GENXAI): a survey, conceptualization, and research agenda. Artif Intell Rev 57(11):289
Shin D, He S, Lee GM et al (2020) Enhancing social media analysis with visual data analytics: a deep learning approach. MIS Q 44(4):1459–1492
Singh S (2023) Facial expression recognition using convolutional neural networks (CNNs) and generative adversarial networks (GANs) for data augmentation and image generation. PhD thesis, University of Nevada, Las Vegas
Sinha M, Li Y, Chung WS et al (2020) Bias correction for supervised learning in email marketing. In: SIGIR eCom'20
Sixt L, Wild B, Landgraf T (2018) RenderGAN: generating realistic labeled data. Front Robot AI 5:66
Speith T (2022) A review of taxonomies of explainable artificial intelligence (XAI) methods. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, pp 2239–2250
Srivastava N, Hinton G, Krizhevsky A et al (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
Stommel W, de Rijk L (2021) Ethical approval: none sought. How discourse analysts report ethical issues around publicly available online data. Res Ethics 17(3):275–297
Sukumar A, Desai A, Singhal P et al (2024) Training against disguises: addressing and mitigating bias in facial emotion recognition with synthetic data. In: 2024 IEEE 18th international conference on automatic face and gesture recognition (FG). IEEE, pp 1–6
Sun Z, Bai J, Wang P et al (2023) Combining deep subspace feature representation based IKPCANet and jointly constraint multi-dictionary learning for facial expression recognition. Artif Intell Rev 56(Suppl 1):937–958
Van der Schyff K, Flowerday S, Furnell S (2020) Duplicitous social media and data surveillance: an evaluation of privacy risk. Comput Secur 94:101822
Voigt P, Von dem Bussche A (2017) The EU general data protection regulation (GDPR). A practical guide, 1st edn, 10(3152676): 10-5555. Springer, Cham
Wu Z, Lischinski D, Shechtman E (2021) StyleSpace analysis: disentangled controls for StyleGAN image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12863–12872
Yan Y, Huang Y, Chen S et al (2019) Joint deep learning of facial expression synthesis and recognition. IEEE Trans Multimedia 22(11):2792–2807
Yuan R, Wang B, Sun Y et al (2022) Conditional style-based generative adversarial networks for renewable scenario generation. IEEE Trans Power Syst 38(2):1281–1296
Zhao G, Huang X, Taini M et al (2011) Facial expression recognition from near-infrared videos. Image Vis Comput 29(9):607–619
Zhu X, Ye S, Zhao L et al (2021) Hybrid attention cascade network for facial expression recognition. Sensors 21(6):2003