

Thermal Emotion Recognition with and without Synthetic Facial Images using Attention-Based EfficientNet5

Tahir Aman

Department of Computer Science and Engineering, Adama Science and Technology University, Ethiopia
E-mail: amantahir21@gmail.com
ORCID iD: <https://orcid.org/0009-0008-3688-8095>

Sintayehu Hirpassa

Department of Computer Science and Engineering, Adama Science and Technology University, Ethiopia
E-mail: sintsha2002@gmail.com
ORCID iD: <https://orcid.org/0000-0003-1129-5214>

Kibreab Adane*

Faculty of Computing and Software Engineering, Arba Minch University, Ethiopia
E-mail: kibreab.adane@amu.edu.et
ORCID iD: <https://orcid.org/0000-0002-3021-5059>

*Corresponding Author

Received: April 29, 2026; Revised: June 8, 2026; Accepted: July 5, 2026; Published: August 8, 2026

Abstract: Humans use emotions to express their feelings and effectively interact with others. Humans express emotions through hands, voice, gestures, and, most importantly, facial expressions. Facial emotion recognition is widely used in human-computer interaction, security, and healthcare. Traditional facial emotion recognition using visible light is often affected by changing lighting conditions. Thermal images could be used as an alternative solution because they rely on physiological heat patterns that remain consistent regardless of illumination. However, the use of thermal images for emotion recognition has not been extensively explored, primarily due to the scarcity of thermal image datasets. The study utilized the recently published Thermal Emotion dataset, which contains 2,250 thermal images before augmentation. After generating a synthetic dataset using cGAN, the datasets expanded to 6,823 images. To prevent dataset leakage, the study used 80% of the data for training, 10% for validation, and 10% for testing. The study used a bilateral filter to reduce noise while keeping relevant edge information, used CLAHE to enhance local contrast in low-intensity regions, making subtle thermal gradients more distinguishable, used Gaussian smoothing to reduce high-frequency noise, resulting in more stable feature extraction, and used attention mechanisms, CBAM, to allow the model to focus on emotion-relevant facial regions through its channel and spatial attention components. The combination of these techniques improved inter-class separability and contributed to measurable gains in recognition accuracy. The study findings show that inclusions of the bilateral filter, CLAHE cGAN, EfficientNetB5, and the CBAM attention mechanism significantly improved accuracy from 97.04% to 98.81%. For each experiment, ResNet18 classifies human facial emotions into five expressions: Happy, Sad, Angry, Natural, and Surprise. These findings imply that synthetic data generation using a cGAN to overcome data scarcity and an attention mechanism provides robust, lightning-independent solutions for facial emotion recognition.

Index Terms: Thermal Image, Facial Emotion Recognition, cGAN, Face Recognition, EfficientNetB5, CBAM

1. Introduction

Emotion represents a feeling-dominated state that affects behavioral reactions as well as psychological and physical alterations [1, 2]. Humans express their emotions through various means, including the movement of hands, pitch of voice, gestures, and above all, through facial expressions, which account for 55% of emotional communication [1]. The human face provides crucial information on emotional states and behaviors, as facial expressions are essential for understanding others' motives during social interactions. Among the diverse techniques in this domain, Facial Emotion Recognition has

demonstrated impressive results [2]. With FER aiding in the interpretation of others' feelings and intentions, it has high relevance in many areas, including healthcare, education, workplace environments, robotics, communication, and psychology, among others [3].

Facial responses are often good predictors of the health or emotional state of the person and are typically best categorized among one of the seven universal emotions: aggression, anxiety, astonishment, sadness, hatred, joy, or indifference [4]. There are significant limitations associated with facial emotion recognition through visible images, including the ease with which individuals can feign expressions and the considerable influence of lighting conditions on image quality [5]. While numerous studies have been conducted to recognize facial expressions from images captured with standard visible light cameras, there is still much to explore in this area [6]. Classifying facial expressions can be challenging due to everyday factors such as shadows, reflections, and low-light conditions. The use of thermal images can mitigate these challenges, as they focus on the thermal distribution in facial muscles, leading to improved accuracy in facial expression classification. Thermal imaging relies on the principle of heat radiation emitted from the skin of the face. Thermal images are employed to evaluate both immediate and chronic stress conditions, as these factors significantly impact blood flow. Variations in blood circulation result in changes in facial temperature, and these temperature fluctuations aid in detecting and refining the thermal imaging process [7].

The unique patterns of superficial blood vessels, veins, and tissue structures beneath the skin significantly contribute to these thermal patterns, making them distinctive for each individual. Facial emotion recognition using thermal imaging is unaffected by skin color or facial features. Temperature on the face elevates when there is an external sound, artifact, or event [5]. Emotion recognition based on visible images is often unreliable due to frequent interference by illumination and intentional manipulation of facial expressions [8]. By comparison, thermal imaging provides a more robust solution due to its ability to capture thermal emissions radiating from the human body, making it less prone to variations in exterior lighting and less vulnerable to intentional facial manipulation. Though thermal image-based emotion recognition has numerous advantages, applications have remained largely not investigated due to a shortage of thermal image datasets. Previous studies, such as [9] research, do not fully account for the most relevant facial features; [8] relies upon a small dataset lacking augmentation; [7] presents poor precision rates for certain emotions; and [10] makes use of datasets where respondents only display a narrow range of expressions. [11] lack of preserving fine facial features, relies upon a small number of subjects, and lacks augmentation.

In response to these shortcomings, this research employed preprocessing, conditional Generative Adversarial Networks (cGANs) for synthetic data generation, and integrated EfficientNet with an attention mechanism for feature extraction. Such a methodology expands the thermal image datasets and highlights the most relevant parts in thermal images, and hence improves facial emotion recognition accuracy. This study proposed human face emotion recognition in thermal images using a cGAN and EfficientNet with an attention mechanism. Overall, the study addressed the following research questions:

- How does attention-based EfficientNetB5 perform both with and without Bilateral Filter and CLAHE?
- How does attention-based EfficientNetB5 perform both with and without Bilateral Filter, CLAHE, and cGAN?
- How does attention-based EfficientNetB5 perform both with and without Bilateral Filter, CLAHE, cGAN, and CBAM?

The novel contributions of the study are presented as follows.

• **Synthetic data generation**

To overcome thermal image dataset shortages, increase dataset diversity, reduce class imbalance, and prevent model overfitting, the study employed a cGAN to generate additional synthetic images.

• **Enhanced preprocessing pipeline**

The study used Bilateral Filters to reduce noise while keeping edge information. Used Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance local contrast, making it distinguishable even in low intensity regions. Used Gaussian Blur to suppress high-frequency noise.

• **Attention-Based Architecture (EfficientNet-B5 + CBAM)**

The study employed EfficientNet-B5 as a backbone, which is more powerful than standard ResNet or VGG models often used in previous thermal image-based research works. Integrated EfficientNet-B5 with the Convolutional Block Attention Module (CBAM), which consists of chain attention to focus on the most informative feature maps while ignoring noisy information.

• **Empirical Validations through incremental experiments**

The study conducted four distinct experiments and confirmed that inclusions of the bilateral filter, cGAN, and the CBAM attention mechanism significantly improved accuracy from 97.04% to 98.81%.

2. Related Works

Assiri et al. (2023) [9] proposed Face emotion recognition based on infrared thermal imagery by the use of machine learning and parallelism. It divides a captured frame into four regions to detect Action Regions. They detect the central point by detecting the tip of the nose, which is further used to differentiate other facial regions. A complex posture estimation method is implemented, as facial postures have a great impact on expressions. However, head postures may bring variations in results. Unlike their approach, which suffers from pose variations, our approach targets preprocessing and attention mechanisms to improve robustness without requiring complicated posture estimation.

Kyal et al. (2020) [12] addressed FER in thermal imaging, focusing on face detection, landmark localization, feature extraction, and parallel deep emotion networks. This study highlights that areas such as the forehead, cheeks, nose, lips, and chin are particularly relevant for accurate emotion recognition in thermal infrared images. However, it does not target any specific area in detail, and the dataset assumes relatively stable head positions. Our approach targets preprocessing and attention mechanisms for improving feature extractions and alleviating small dataset problems through the use of cGAN for training data expansion.

Prabhakaran et al. (2021) [8] used thermal images and a pre-trained modified version of ResNet152 for Facial expression recognition. The model relies on a small dataset, and no data augmentation techniques were applied. We overcome this limitation by adding cGAN-based synthetic data augmentation and stochastic augmentation, which increase data diversity and decrease overfitting.

Elbarawy et al. (2019) [10] experimented on Facial Expression Recognition in Thermal Images using Deep Learning Techniques. It takes 90 images with varying rotations, as well as occlusion due to glasses and pose; 60 for training and 30 for testing. Only the poses with less than 45° rotation were selected. The subject has only three different expressions: Surprise, Happy, and Angry. We alleviate small-dataset problems through the use of cGAN for training data expansion with five-class emotion recognition.

Prasad et al. (2023) [11] presented a four-stage approach to recognizing human faces in thermal images: pre-processing, feature extraction, detection, and classification. In the pre-processing stage, input thermal face images are cropped by the Difference of the Gaussian filter and normalized by a median filter to handle variations in lighting conditions. Then, the facial features' position, shape, and object presence will be extracted using the EfficientNet approach. After that, face detection is performed using YOLOv4, followed by the classification of emotions by DenseNet from the thermal images. However, it lacks fine facial features, relies upon a small number of subjects, and lacks augmentation. We extend this approach by employing Bilateral Filter, Gaussian Blur, and CLAHE, and integrate EfficientNet-B5 with CBAM for enhanced feature extraction, and cGAN for set expansion for dealing with scalability.

Chaitanya et al. (2020) [7] used YOLO methodology and a Darknet framework for predicting human emotions. Background elimination, pre-processing, a training phase, and a test phase are included in the work. The model is good at detecting happiness, sadness, surprise, and disgust. Anger, fear, and neutral are some of those aspects where the model had less accurate precision scores. By incorporating EfficientNet-B5 with CBAM, our approach improves discriminative feature extraction and aids in improving emotion recognition.

Previous studies, such as those by [9, 12], achieved high precision; however, their requirements for specific face areas and static head positions. Likewise, [8, 10] illustrated the capability of deep learning using ResNet and other models; however, these studies were limited by the use of highly limited datasets, as well as the lack of data augmentation. Further advanced methods, such as those by [11], employed the use of EfficientNet with DenseNet classifiers; however, their requirements for limited subjects, lack of preserving fine facial features, lack of enhanced image quality, and lack of GAN-based augmentation limited them for scaling. Differing from these deficiencies, this paper makes up for them by employing cGAN-based data generation for mitigating the scarcity of datasets, integrating EfficientNet-B5 and Convolutional Block Attention Module for enhancing feature learning, and a ResNet-18 classifier for effective classification.

3. Methodology

3.1. Research Design

As a vital biometric identification technology, face recognition is widely employed in identity verification and security surveillance mechanisms. Current research mainly depends on high-definition RGB camera-based approaches. However, these approaches are prone to several factors, such as lighting conditions and disguises [15]. To address these concerns, this study focuses on human face emotion recognition in thermal images based on the utilization of preprocessing, data augmentation, synthetic data generation, feature extraction, and classification. Our study introduces a comprehensive pipeline designed to enhance facial emotion recognition accuracy using a thermal image dataset through four main stages. I) At the data preprocessing stage, the study uses Bilateral Filtering to reduce noise while preserving sharp facial edges, uses CLAHE to redistribute pixel intensities, and highlights fine texture patterns. II) To overcome dataset scarcity, the study uses a cGAN to generate high-quality thermal images, which increases the diversity of training data and reducing risk of overfitting. III) For advanced feature extraction, the study uses the combination of EfficientNet-B5 and CBAM. The EfficientNet-B5 is used as a backbone due to its efficiency in balancing network width and depth. It

is integrated with CBAM, which uses channel and spatial attention to locate the network in the most useful regions for emotion, such as the mouth, eyes, and forehead. IV) For robust classifications, the extracted features are passed to a ResNet-18 classifier. This architecture leverages residual learning to keep stable training and sharply map complex thermal features into five emotional classes: Happy, Sad, Angry, Natural, and Surprise. The proposed architecture of the model is shown in Figure 1. Compared to previous studies, our approach ensures promising results, as shown in Table 1.

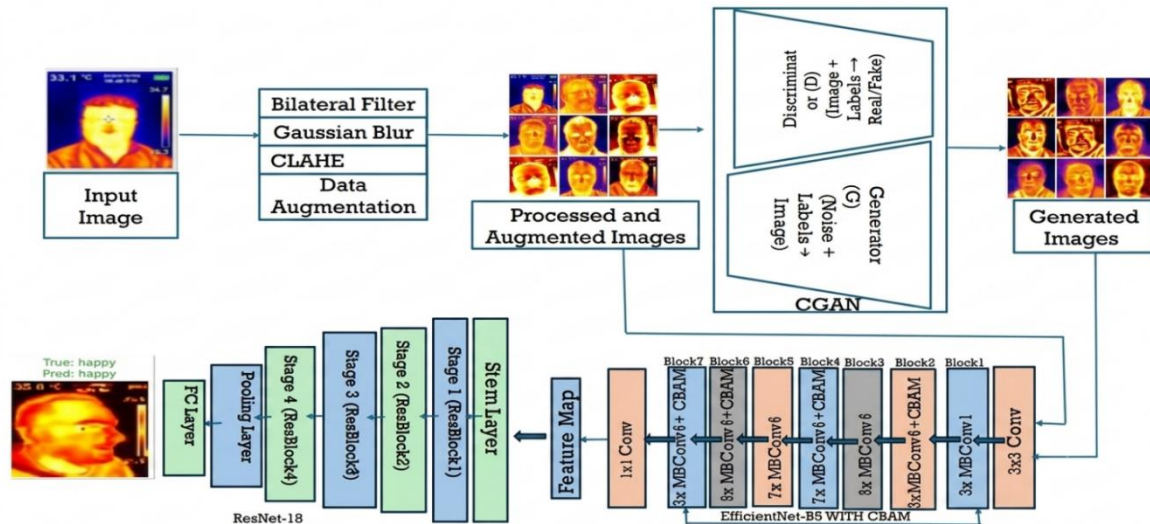


Fig. 1. Proposed Model architecture

Table 1. Comparison of related studies on thermal emotion recognition, including their potential limitations

Authors	Research focus	Methodology	Dataset used	Accuracy	Limitations
Assiri et al. [9]	“Face Emotion Recognition Based on Infrared Thermal Imagery by Applying Machine Learning and Parallelism”	Image Segmentation, Image Registration and Preprocessing, Feature Extraction and Classification Using CNN, a machine learning technique and parallelism.	CK+ dataset, JAFFE dataset	96.87 % , 91.8%	Only four active regions are selected to create the training and testing datasets. The nose tip serves as a reference point to identify and differentiate other regions. However, variations in the head may introduce inconsistencies.
Kyal et al. [12]	“Human Emotion Recognition from Spontaneous Thermal Image Sequence Using GPU Accelerated Emotion Landmark Localization and Parallel Deep Emotion Net”	The thermal domain based on face detection, facial landmark localization, feature extraction, and parallel deep emotion network	NVIE Dataset	94%	The analysis showed that facial regions like the forehead, left and right cheeks, nose, lips, and chin are more reliable for thermal infrared facial emotion recognition. However, it does not target any specific area in detail, and the dataset assumes relatively stable head positions.
Chaitanya et al. [7]	“Human Emotions Recognition from Thermal Images Using YOLO Algorithm”	Background elimination, preprocessing, training phase, and test phase using Yolo Algorithm	NVIE Dataset	65 %	The model performs well in detecting emotions such as happiness, sadness, surprise, and disgust. However, it shows lower precision scores for emotions like anger, fear, and neutral.
Prabhakaran et al. [8]	“Thermal Facial Expression Recognition Using Modified ResNet152”	Data Preprocessing, Training, and Testing using a modified ResNet152 model.	NVIE Dataset	87.3%	The model relies on a small dataset, and no data augmentation techniques were applied.

Elbarawy et al. [10]	“Facial Expressions Recognition in Thermal Images based on Deep Learning Techniques”	Facial Expressions Recognition in Thermal Images Based on Deep Learning Techniques	IRIS dataset	96.7%	The system utilizes 90 images featuring variations in rotation, occlusion by glasses, and different poses. Only poses with rotations of less than 45° were included. The subject displays three expressions: surprise, happiness, and anger.
Prasad et al. [11]	“Human Face Emotions Recognition from Thermal Images Using DenseNet”	Preprocessing, feature extraction, detection, and classification of the emotions on faces in thermal images, using the DenseNet technique.	RGB-D-T dataset	95.97%	Only 51 subjects are included, but the absence of GAN-based or other augmentation methods
Ours (Proposed)	“Thermal Emotion Recognition with and without Synthetic Facial Images using Attention-Based EfficientNet5”	Combined Bilateral Filter, CLAHE, cGAN, and EfficientNet-B5 with CBAM, ResNet18 Combined Bilateral Filter, CLAHE, and EfficientNet-B5, ResNet18 Combined Bilateral Filter, CLAHE, cGAN, EfficientNet-B5, and ResNet18	Comprehensive Facial Thermal Dataset [13]	98.81% 97.21% 98.01%	Our study didn’t conduct computational efficiency analysis considering inference time, FLOPs, memory usage, or energy consumption.

3.2. Dataset Information and Augmentation Process

The original dataset was created by [13] and published in March 2024, contains 2250 thermal images obtained via a UNI-T UTi165A camera and is primarily used for face detection, recognition, and emotion analysis purposes. Each dataset is represented by 225 facial recognition images and 450 emotion-specific images, namely Sad, Angry, Happy, Natural, and Surprise expressions. It incorporates diverse conditions such as shooting angles, color palettes, and zoom levels. It offers -10°C to 400°C temperature and 19,200 pixels resolution. It is a vital asset for the improvement of thermal imaging technology, enabling algorithmic development and benchmarking for facial recognition across diverse palettes. It has significant contributions in nurturing interdisciplinary collaborations across the domains of computer vision, neuroscience, and psychology [13].

We applied the Bilateral Filter to enhance contrast and reduce noise; Gaussian Blur to smooth the image; CLAHE to contrast in low-light conditions, normalizing the images to standardize intensity levels, and resizing them to ensure uniform dimensions; and used data augmentation techniques, cGAN to address data scarcity and class imbalance. The Augmentation process includes flipping, rotation, shifts, and scaling for robustness.

Due to the limited dataset size, a cGAN was trained to generate synthetic images for each emotion class. The cGAN combines the generator and discriminator into one system to train them together. In the training process of the cGAN, the generator is responsible for creating fake images by combining random noise vectors with class labels, such as different facial emotions. These synthetic images are then passed to the discriminator, which evaluates them and produces a prediction indicating whether the image is real or fake. Training was monitored using generator and discriminator losses. Generated images were evaluated using the Frechet Inception Distance, confirming high similarity to real thermal data. The synthetic dataset increased class diversity and reduced overfitting.

After generating a synthetic dataset using a conditional GAN, the whole set of datasets includes 6,823 images divided into three independent sets: 2,054 images for subject identification tasks, 2,284 images for face detection, and 2,485 images for emotion recognition. Datasets include variations in distance, pose, and facial temperature distribution. To prevent dataset leakage, this study used 80% of the data for training, 10% for validation, and 10% for testing.

The Generator architecture formed from the latent vector and class labels as input and generated a thermal image (1 channel, grayscale) as output. The embedding layer in the generator architectures is used to convert class labels into embeddings of size $\text{embed_dim} = 10$. The latent vector and label embeddings are concatenated along the channel dimension. The generator architecture is shown in Table 2.

The discriminator network accepts a gray-scale image (1 channel) and class labels and produces a probability score, with 0 being fake and 1 being real. To condition on class information, the labels are first fed through an embedding layer that transforms them into 50-sized embeddings. The embeddings are then reshaped and expanded using $\text{label_emb.view(labels.size(0), -1, 1, 1)}$ to have the same spatial resolution as the image, after which they are concatenated with the image along the channel axis. Therefore, the first convolutional block is presented with a total of 1+50 input channels. The discriminator consists of five convolutional blocks followed by a sigmoid activation, which provides the probability that the input image is real or fake. The discriminator architecture is shown in Table 2.

Table 2. Discriminator Architecture

Layer Type	Details
Input Layer	Conv2d(in_channels=img_channels + 50, out_channels=64, kernel_size=4, stride=2, padding=1)LeakyReLU(negative_slope=0.2)
Hidden Layer 1	Conv2d(64, 128, kernel_size=4, stride=2, padding=1) BatchNorm2d(128) LeakyReLU(0.2)
Hidden Layer 2	Conv2d(128, 256, kernel_size=4, stride=2, padding=1) BatchNorm2d(256) LeakyReLU(0.2)
Hidden Layer 3	Conv2d(256, 512, kernel_size=4, stride=2, padding=1) BatchNorm2d(512) LeakyReLU(0.2)
Output Layer	Conv2d(512, 1, kernel_size=4, stride=1, padding=0) Sigmoid ()

After preprocessing the images, we then extracted the features by using EfficientNet-B5 with an attention mechanism. EfficientNet-B5 uses compound model scaling to balance network depth, width, and resolution. It is well-suited for resource-efficient but high-accuracy feature extraction. To further improve the feature representation learned by EfficientNet-B5, Convolutional Block Attention Modules are used on 2, 4, 6, and 7 blocks of the network. These attention modules operate on the features both spatially and channel-wise, making the model focus on more informative regions while ignoring less relevant or noisy information.

The input layer of EfficientNet-B5 takes images of size $(456 \times 456 \times 3)$. EfficientNet-B5 is used in combination with the Convolutional Block Attention Module to extract deep feature representations, while the default classifier is discarded. The EfficientNet-B5 with CBAM MBConv Blocks architecture is shown in Table 3.

Table 3. EfficientNet-B5 with CBAM MBConv Blocks

Layer Type	Details
Input Layer	Image Input: $(456 \times 456 \times 3)$
Stem Convolution	Conv2d(3, 48, kernel_size=3, stride=2, padding=1) $\rightarrow$ BatchNorm2d $\rightarrow$ Swish
MBConv Stage 1	$3 \times$ MBConv1 (expansion=1, kernel=3, output=24, stride=1)
MBConv Stage 2	$5 \times$ MBConv6 (expansion=6, kernel=3, output=40, stride=2)+ CBAM after block 7
MBConv Stage 3	$5 \times$ MBConv6 (expansion=6, kernel=5, output=64, stride=2)
MBConv Stage 4	$7 \times$ MBConv6 (expansion=6, kernel=3, output=128, stride=2)+ CBAM after block 19
MBConv Stage 5	$7 \times$ MBConv6 (expansion=6, kernel=5, output=176, stride=1)
MBConv Stage 6	$9 \times$ MBConv6 (expansion=6, kernel=5, output=304, stride=2)+ CBAM after block 35
MBConv Stage 7	$3 \times$ MBConv6 (expansion=6, kernel=3, output=512, stride=1)+ CBAM after block 38
Final Convolution	Conv2d(512, 2048, kernel_size=1, stride=1) $\rightarrow$ BatchNorm2d $\rightarrow$ Swish
Pooling Layer	Global Average Pooling (GAP)
Dropout Layer	Dropout(p=0.4)

Finally, a fully connected classification layer maps the learned representation to the five intended emotion classes: happiness, sadness, anger, surprise, and neutrality. Thus, the forward propagation follows the scheme: input projection $\rightarrow$ residual blocks (according to the ResNet-18 architecture) $\rightarrow$ global average pooling $\rightarrow$ final fully connected layer $\rightarrow$ logits. With the use of residual connections, the ResNet-18 head provides stable training and more abundant feature abstraction than a simple multilayer perceptron, with relatively modest computational requirements. The ResNet-18 architecture is shown in Table 4.

Table 4. ResNet-18 architecture

Stage / Layer	Details	Output Shape
Input	Image $(224 \times 224 \times 3)$	$224 \times 224 \times 3$
Conv1	7×7 conv, 64 filters, stride=2	$112 \times 112 \times 64$
MaxPool	3×3 max pool, stride=2	$56 \times 56 \times 64$
Stage 1 (ResBlock1)	$2 \times$ BasicBlock, 64 filters, stride=1	$56 \times 56 \times 64$
Stage 2 (ResBlock2)	$2 \times$ BasicBlock, 128 filters, stride=2 in first block	$28 \times 28 \times 128$
Stage 3 (ResBlock3)	$2 \times$ BasicBlock, 256 filters, stride=2 in first block	$14 \times 14 \times 256$
Stage 4 (ResBlock4)	$2 \times$ BasicBlock, 512 filters, stride=2 in first block	$7 \times 7 \times 512$
Head (Pooling)	Global Average Pooling (GAP)	$1 \times 1 \times 512$
Head (FC Layer)	Fully Connected (512 $\rightarrow$ num_classes), Softmax / Logits	Five_classes

3.3. Hyperparameters for Image Generation and Model Training

Table 5 summarizes different experiments which are conducted in this study. In the first experiment, a baseline was trained using the original datasets to provide a reference for the later comparison with the proposed model. In the second experiment, bilateral filtering and CLAHE preprocessing were applied, with EfficientNet-B5 used for feature extraction. The goal of this experiment is to address the impact of preprocessing on thermal facial emotion recognition. In the third experiment, bilateral filter, CLHE, and cGAN -generated synthetic data were combined with EfficientNet-B5 without attention to investigate the impact of synthetic data on model performance. Finally, the fourth experiment was conducted by utilizing a bilateral filter, CLAHE, cGAN, and EfficientNet-B5 integrated with CBAM to address the impact of attention-based feature extraction on classification performance. The cGAN was trained independently before the classification model to generate emotion-conditioned synthetic thermal images. Due to the instability commonly associated with adversarial training, a separate hyperparameter optimization process was conducted using a greedy search strategy. During greedy search, one hyperparameter was adjusted at a time while others were fixed.

The input images are also resized to a standard size of (224, 224) to ensure consistency in the dataset. The pixel values are normalized with a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225], which is the default for models pre-trained on ImageNet. The model is trained at a batch size of 64 to enable data processing in batches. Shuffling is permitted during training to ensure that each batch is randomly selected from the datasets, which favors better generalization. Shuffling is disabled for validation and test phases to maintain the consistency of data order. Data loading is made efficient by setting the number of workers to 4, permitting multi-threaded data loading for faster batch fetching, which contributes to training efficiency. The hyperparameters used in image generation are listed in Table 5, and the hyperparameters used to train the model are listed in Table 6.

Table 5. Hyperparameters for generating Images

Parameter	Explored Range	Final Selected Value
Latent vector dimension (z_dim)	512-2048	2048
Embedding dimension (embed_dim)	10	10
Number of classes	5	5
Image resolution	64×64	64×64
Generator architecture	5 ConvTranspose2d layers with BatchNorm + ReLU	5 ConvTranspose2d layers with BatchNorm + ReLU
Discriminator architecture	5 Conv2d layers with BatchNorm + LeakyReLU	5 Conv2d layers with BatchNorm + LeakyReLU
Loss function	Binary Cross-Entropy Loss	Binary Cross-Entropy Loss
Optimizer	Adam	Adam
Learning rate	$1.6e-4 - 2e-4$	0.00016
Adam betas	(0.5, 0.999)	(0.5, 0.999)
Batch size	16 - 64	64
Number of epochs	100 - 400	400

Table 6. Hyperparameters for Model Training

Hyper parameter	Explored Range	Final Selected Value
Loss Function	CrossEntropyLoss	CrossEntropyLoss
Optimizer	Adam	Adam
Learning Rate	$1e-4 - 2e-4$	$1e-4$
Weight Decay	$1e-4 - 2e-4$	$1e-4$
Batch Size	8, 16, 32	16
Epochs	20 - 50	50
Early Stopping Patience	5 - 10	10
Dropout rates	0.3 - 0.5	0.4

3.4. Experimental Setup

The study used the following software and hardware specifications. PyTorch was used to build and train models, while Tensorboard was used as a visualization tool to monitor, debug, and optimize those models, namely EfficientNet-B5 and ResNet-18. Anaconda version 23.7.4 was used as a platform that manages programming environments and software packages, while Jupyter Notebook served as an interactive, browser-based application bundled with Anaconda.

Google Colab was used to provide access to free cloud-based GPUs, significantly accelerating the training of computationally intensive deep learning models. Python library packages, including NumPy, pandas, SciPy, Matplotlib, and scikit-learn, were used to cover the entire data pipeline: from raw mathematical computing to data cleaning, advanced scientific calculations, visual presentation, and predictive modeling. Table 7 illustrates the hardware specifications used for experiments.

Table 7. Hardware Tools Specification

Hardware Component	Specification
Processor	Intel(R) Core (TM) i7-8700M CPU @ 3.20GHz 3.19 GHz
Installed RAM	8.00 GB
Operating System	Windows 10
System Type	64-bit operating system, x64-based processor
Graphics	Intel ® HD Graphics 520 / NVIDIA Corporation
GPU	NVIDIA GeForce RTX 2070 SUPER
Driver Version	560.94
CUDA Version	12.6
GPU Memory	8192 MB (8 GB)

The evaluation of the performance was conducted in terms of standard classification measures, such as accuracy, precision, recall, F1-score, and AUC-ROC.

4. Results and Discussion

The performance of the cGAN was quantified based on the implementation of Frechet Inception Distance stability (FIDs), which is a known metric for assessing the quality and realism of the synthesized images. Within the framework of the classification task, we have used various standard model evaluation metrics presented below.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively.

4.1. Experimental Results on Data Augmentation(cGAN)

To train cGAN, an optimal configuration was selected based on generator–discriminator loss balance, visual quality of generated images, and Frechet Inception Distance stability. In our experimentation, the final hyperparameter configurations were selected using a greedy search strategy based on validation accuracy for the classification model and training stability for the cGAN. The selected configurations provided the optimal balance between convergence speed, model generalization, and computational efficiency. The final selected configuration ensured stable adversarial training and high-quality emotion-conditioned thermal image generation.

The CrossEntropyLoss function acts as the loss function, appropriate for multi-class classification tasks. Optimizathe Adam optimizer, where the learning rate has been set at 0.0001, to provide a balance between convergence rate and stability. Training is performed for 50 epochs, and early stopping is activated to monitor validation performance and terminate training if there is no improvement in 10 consecutive epochs, thus preventing overfitting. For weight initialization, the model utilizes pre-trained EfficientNet-B5 weights, thus facilitating transfer learning of a model trained on the large ImageNet dataset.

To measure the performance of the cGAN in generating synthetic thermal facial emotion images, we tracked the generator and discriminator loss curves, which are shown in Figure 2. Figure 3 illustrates sample-generated images.

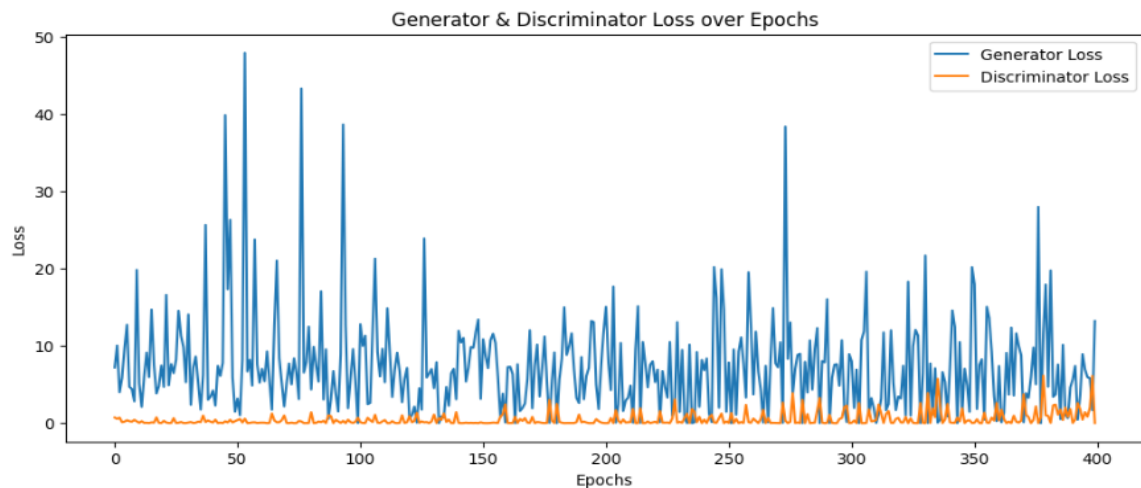


Fig. 2. cGAN generator and discriminator loss

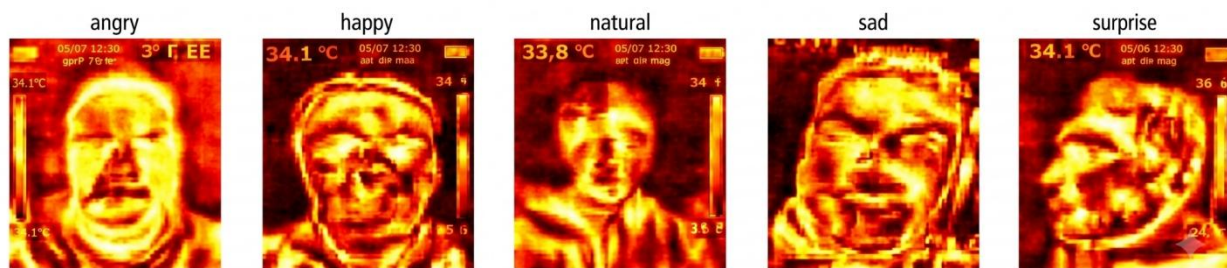


Fig. 3. Sample generated images

Realism and generated image quality were measured quantitatively using the Fréchet Inception Distance Stability (FIDs). A FID score below 30 indicates strong similarity between the generated and real distributions in thermal datasets. Decreasing values for FID show increasing similarity to actual image distributions. During our experiments, the calculated FID value of 20.7062 confirmed that the generated thermal images closely matched the distribution of real samples while maintaining diversity, which is shown in Figure 4.

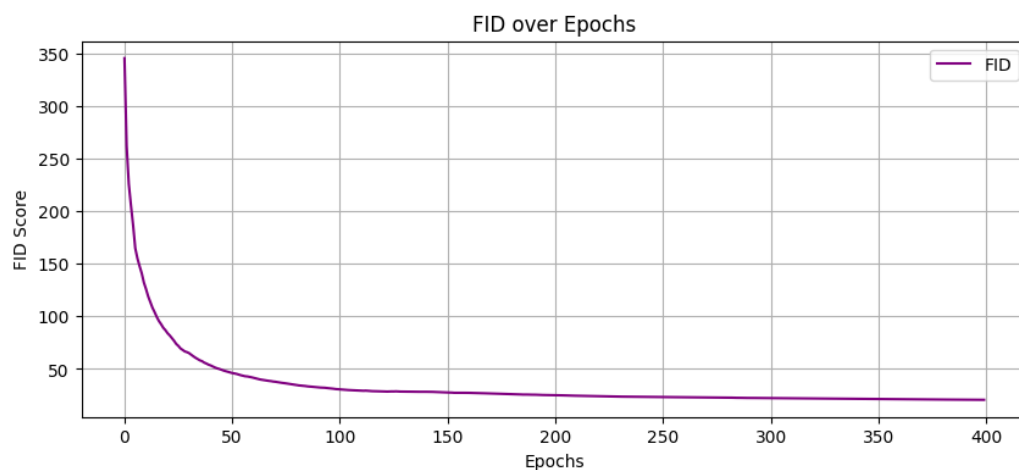


Fig. 4. FID score based on epoch

To systematically evaluate the effectiveness of each component in our proposed approach. The performance of each configuration was assessed using accuracy, precision, recall, and F1-score metrics.

4.2. Attention-based EfficientNetB5 performance with and without Bilateral Filter and CLAHE

We conducted a series of experiments with incremental model configurations. This allowed us to isolate the contribution of each element, namely, data augmentation using cGAN, feature extraction using EfficientNet-B5, and CBAM. In all experiments, ResNet-18 served as the classification model, leveraging its residual connections to improve training stability and enhance the mapping from high-dimensional features to emotion labels. The baseline model was trained using the original comprehensive facial thermal dataset without synthetically generated data and an attention mechanism. The model achieved a test accuracy of 97.04%. This setup was utilized as a baseline to validate the impact of the enhancements made in subsequent experiments. Figure 5 illustrates the Training and Validation graph of the baseline model (Without Data Augmentation Techniques).

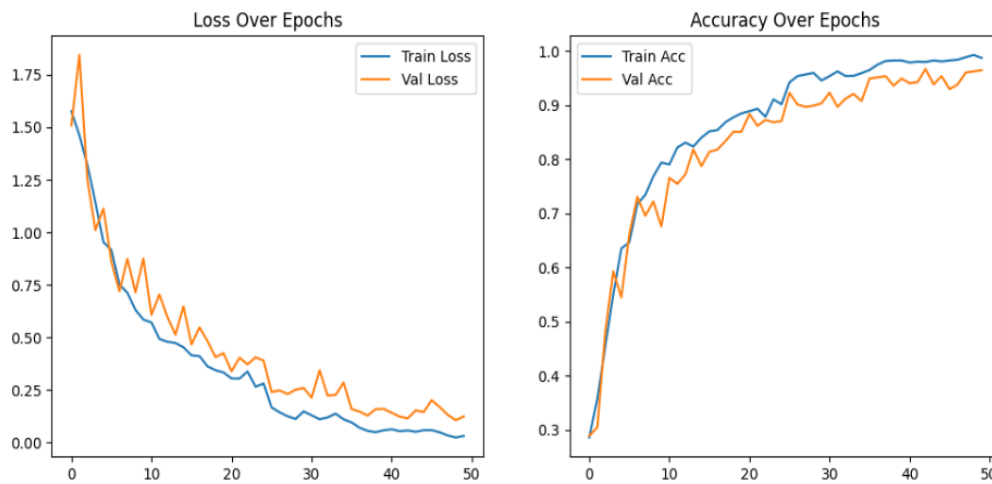


Fig.5. Training and Validation graph of baseline model (Without Data Augmentation Techniques)

Table 8 shows the baseline models' class-wise performance measures based on Precision, recall, and F1-score. Surprise face emotion class attained the top performance by scoring 99% average precision, recall, and F1-score, while the Happy face emotion class ranked second by scoring 98% average precision, recall, and F1-score.

Table 8. Baseline model Performance per Face Emotion Class

Face Emotion class	Precision	Recall	F1-Score	Average
Happy	98%	97%	98%	98%
Natural	95%	98%	96%	96%
Surprise	99%	99%	99%	99%
Sad	99%	95%	97%	97%
Angry	95%	96%	96%	96%

In the second configuration, we combined the Bilateral Filter, CLAHE, Efficient Net B5, and used ResNet18 as a face emotion classifier. After adding the Bilateral Filter and CLAHE, the model achieved a test accuracy of 97.21%, which was a slight improvement over the baseline model accuracy by 0.17%, as shown in Table 10. Happy and Sad face emotion classes attained the top performance by scoring 98% average precision, recall, and F1-score, while Angry and Surprise face emotion classes ranked second by scoring 97% average precision, recall, and F1-score, as shown in Table 9.

Table 9. Performance Metrics per Emotion Class after combining the Bilateral Filter, CLAHE, Efficient Net B5, and ResNet18

Face Emotion	Precision	Recall	F1-Score	Average
Happy	99 %	98%	98%	98%
Natural	94%	98%	96%	96%
Angry	99%	95%	97%	97%
Surprise	95%	100%	97%	97%
Sad	100%	96%	98%	98%

Table 10. EfficientNetB5 performance with and without Bilateral Filter and CLAHE

Models	Accuracy	Precision	Recall	F1 score
Baseline- EfficientNet	97.04%	97%	97%	97%
Combined Bilateral Filter, CLAHE, and EfficientNet-B5	97.21%	97%	97%	97%

Figure 6 illustrates the Training and Validation graph of the combined Bilateral Filter, CLAHE, EfficientNet-B5, and ResNet 18.

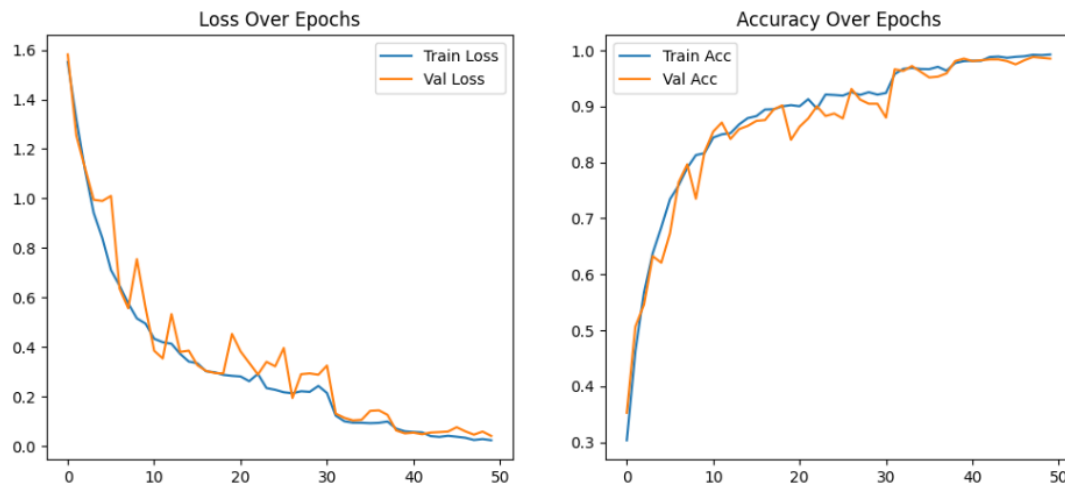


Fig. 6. Training and Validation graph of the combined Bilateral Filter, CLAHE, EfficientNet-B5, and ResNet 18

4.3. Attention-based EfficientNetB5 performance with and without Bilateral Filter, CLAHE, and cGAN

In the third configuration, we combined the Bilateral Filter, CLAHE, cGAN, EfficientNet B5, and used ResNet18 as a face emotion classifier. As shown in Table 12, following the addition of cGAN to increase data diversity through synthetic generation, the model achieved a test accuracy of 98.01%, which was a 0.97% improvement over the baseline model accuracy. Happy and Sad face emotion classes attained the top performance by scoring 99% average precision, recall, and F1-score, while Angry and Surprise face emotion classes ranked second by scoring 98% average precision, recall, and F1-score, as shown in Table 11.

Table 11. Performance Metrics per Emotion Class after combining the Bilateral Filter, CLAHE, cGAN, Efficient Net B5, and ResNet18

Face Emotion	Precision	Recall	F1-Score	Average
Happy	98%	100%	99%	99%
Natural	99%	94%	96%	96%
Angry	97%	99%	98%	98%
Surprise	96%	99%	98%	98%
Sad	99%	99%	99%	99%

Table 12. EfficientNetB5 performance with and without Bilateral Filter, CLAHE, and cGAN

Models	Accuracy	Precision	Recall	F1 score
Baseline- EfficientNet	97.04%	97%	97%	97%
Combined Bilateral Filter, CLAHE, cGAN, and EfficientNet-B5	98.01%	98%	98%	98%

Figure 7 illustrates the Training and Validation graph of the combined Bilateral Filter, CLAHE, cGAN, EfficientNet-B5, and ResNet 18.

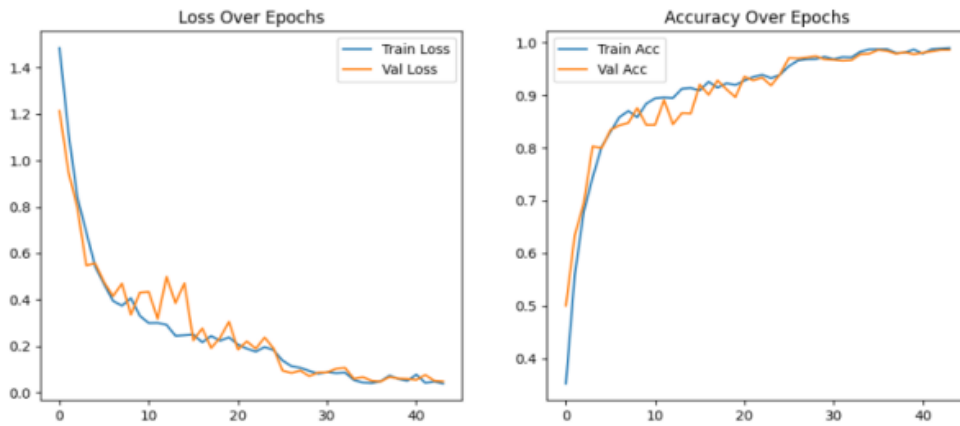


Fig. 7. Training and Validation graph of the combined Bilateral Filter, CLAHE, cGAN, EfficientNet-B5, and ResNet 18

4.4. Attention-based EfficientNetB5 performance with and without Bilateral Filter, CLAHE, cGAN, and CBAM

The fourth configuration incorporated the Bilateral Filter, CLAHE, cGAN, and EfficientNet-B5, which was integrated with CBAM to enhance feature extraction. The ResNet18 was used as a face emotion classifier. As shown in Table 14, incorporating CBAM into EfficientNet-B5 markedly improved performance, achieving a test accuracy of 98.81%, which was a 1.77% improvement over the baseline model accuracy. This result confirmed the role of selective channel weighting and spatial attention in enhancing feature representation and enabling the model to focus on the most informative region for emotion classification. Natural face emotion class attained a perfect (100%) average precision, recall, and F1-score, while the Sad face emotion class ranked second by scoring 99% average precision, recall, and F1-score, as shown in Table 13.

Table 13. Performance Metrics per Emotion Class after combining the Bilateral Filter, CLAHE, cGAN, Efficient Net B5 with CBAM, and ResNet18

Face Emotion	Precision	Recall	F1-Score	Average
Surprise	98%	98%	98%	98%
Sad	100%	99%	99%	99%
Angry	99%	98%	98%	98%
Natural	99%	100%	100%	100%
Happy	98%	99%	98%	98%

Table 14. EfficientNetB5 performance with and without Bilateral Filter, CLAHE, cGAN, and EfficientNet-B5 with CBAM

Models	Accuracy	Precision	Recall	F1 score
Baseline- EfficientNet	97.04%	97%	97%	97%
Combined Bilateral Filter, CLAHE, cGAN, and EfficientNet-B5 with CBAM	98.81%	99%	99%	99%

Figure 8 illustrates the Training and Validation graph of the combined Bilateral Filter, CLAHE, cGAN, EfficientNet-B5 with CBAM, and ResNet 18.

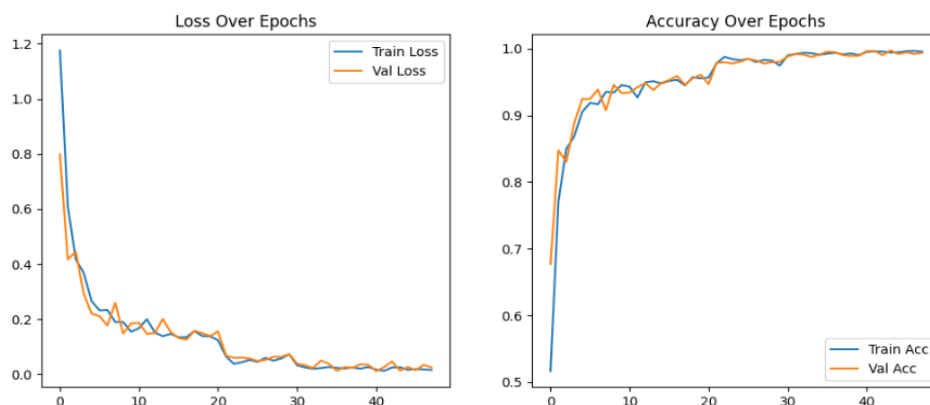


Fig. 8. Training and Validation graph of the combined Bilateral Filter, CLAHE, cGAN, EfficientNet-B5 with CBAM, and ResNet 18

5. Discussion

Thermal images are often affected by low contrast and sensor noise, which can obscure subtle temperature variations associated with emotional states. The bilateral filter effectively reduced noise while preserving edge information, preventing the loss of important facial thermal contours. CLAHE enhanced local contrast in low-intensity regions, making subtle thermal gradients more distinguishable. Additionally, Gaussian smoothing helped reduce high-frequency noise, resulting in more stable feature extraction. The combination of these techniques improved inter-class separability and contributed to measurable gains in recognition accuracy. One of the major challenges in thermal emotion recognition is limited dataset availability. To address this limitation, a conditional GAN was employed to generate synthetic thermal images conditioned on emotion labels. The generated samples increased dataset diversity and reduced class imbalance. Unlike traditional augmentation methods such as rotation or flipping, cGAN-generated images introduced new thermal variations, enabling the model to generalize better to unseen samples. The inclusion of synthetic data reduced overfitting and improved validation performance, demonstrating the effectiveness of adversarial data augmentation in low-resource thermal imaging scenarios.

The choice of EfficientNet-B5 as the backbone network further contributed to improved performance. EfficientNet employs compound scaling, balancing network depth, width, and resolution to maximize accuracy while maintaining parameter efficiency. Thermal facial images contain subtle temperature distributions that require high-resolution feature representation. EfficientNet-B5 provided stronger feature extraction capability compared to conventional architectures such as VGG or standard ResNet models, resulting in improved emotion discrimination. The proposed model takes a preprocessed thermal facial image as input and predicts emotional class among the predefined categories, such as happy, sad, angry, surprise, and neutral. Features are first extracted by the EfficientNet-B5 backbone enhanced with CBAM. These features are then passed through a ResNet-18 classification head, which maps the high-dimensional feature representations to emotion classes. Finally, a SoftMax activation generates a probability distribution across all possible emotional classes, and the class having the maximum probability is identified as the predicted emotion.

Empirical results show that the combination of Bilateral Filter, CLAHE, cGAN, EfficientNet-B5, and ResNet18 increased accuracy from 97.21% to 98.01% before the attention mechanisms like CBAM were applied. The integration of the CBAM with EfficientNet-B5 provided the final boost in performance because CBAM allowed the model to focus on emotion-important facial regions through its channel and spatial attention components. The study conducted four distinct experiments to isolate the value of each component. With Baseline, the original dataset and standard EfficientNet achieved 97.04%. Adding Bilateral Filtering and CLAHE raised the accuracy to 97.21%. Adding cGAN-generated images further increased accuracy to 98.01%. Final Proposed Model, combining all the above with CBAM reached a peak 98.81%. The findings highlighted that the model is robust across all tested emotional states. As shown in Table 13, the Natural face emotion class achieved a perfect average recall, precision, and F1-score of 100%, while face emotion classes such as Sad, Happy, Angry, and Surprise maintained average recall, precision, and F1-score from 98% to 99%. This suggests that the thermal-based feature extraction is highly reliable for distinguishing even similar emotional heat patterns. Table 15 demonstrates the model performance Comparison of each experimental step in terms of accuracy, Precision, Recall, and F1-score.

Table 15. Summary of Experimental Results Comparison

Models	Accuracy	Precision	Recall	F1 score
Baseline- EfficientNet	97.04%	97%	97%	97%
Combined Bilateral Filter, CLAHE, and EfficientNet-B5	97.21%	97%	97%	97%
Combined Bilateral Filter, CLAHE, cGAN, and EfficientNet-B5	98.01%	98%	98%	98%
Combined Bilateral Filter, CLAHE, cGAN, and EfficientNet-B5 with CBAM	98.81%	99%	99%	99%

The key advantage of the AUC-ROC metric is that it measures the effectiveness of models, considering True Positive and False Positive rates, which remains valid regardless of class imbalance or how data classes are distributed, making the model reliable for judging classifier performance in any scenario [14]. Figure 9 illustrates AUC-ROC results of the top-performing model, i.e., Combined Bilateral Filter, CLAHE, cGAN, and EfficientNet-B5 with CBAM. It shows that the proposed model attained more than 99.97% for each class, namely, Angry, Happy, Neutral, Sad, and Surprise. These findings demonstrate that the model is more likely to distinguish between the Positive and Negative classes as a result of scoring an AUC-ROC value closest to 1.

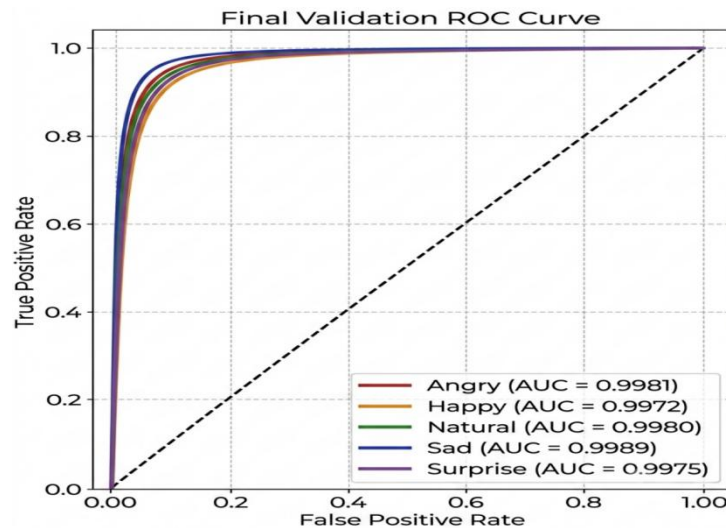


Fig. 9. AUC-ROC results of the Combined Bilateral Filter, CLAHE, cGAN, and EfficientNet-B5 with CBAM

Although our experimental findings have shown potential in advancing human facial emotion detection through thermal imaging, a lot is still available for research and development. Some potential areas of focus for future research are listed as follows.

- This study introduced the primary emotions such as happiness, sadness, anger, neutrality, and surprise. Future research would involve incorporating secondary emotions such as fear, disgust, contempt, and confusion.
- Although thermal imaging is good at recognizing emotions, its combination with other data like visible images, speech, physiological signals, or facial landmarks could enhance the performance. Multimodal studies are worth exploring in the future.
- This study incorporates cGAN and CBAM for improving robustness. Future improvement of robustness can be done by incorporating context models, background subtraction, or handling occlusions and environmental variations for improved feature extraction.
- The future research is expected to employ Explainable AI (XAI) techniques to interpret the contribution of thermal features for emotion classification, enhancing trust and transparency.
- The future research is expected to conduct the models' computational efficiency analysis by including inference time, FLOPs, memory usage, or energy consumption.

6. Summary and Conclusions

Traditional facial emotion recognition using visible light is often affected by changing lighting conditions. Thermal images could be used as an alternative solution because they rely on physiological heat patterns that remain consistent regardless of illumination. However, the field is hindered by a significant shortage of thermal image datasets. To address dataset scarcity, the study employed cGAN to increase dataset diversity, reduce overfitting, and improve the model's ability to generalize to unseen data. The study utilized the recently published Thermal Emotion dataset, which contains 2,250 thermal images before augmentation. After generating a synthetic dataset using cGAN, the datasets expanded to 6,823 images. The preprocessing stage played a crucial role in enhancing feature quality. At the data preprocessing stage, the study uses Bilateral Filtering to reduce noise while keeping sharp facial edges, Gaussian Blur for smoothing, and CLAHE to enhance local contrast and reveal fine texture patterns in low-visibility thermal images. The study conducted four distinct experiments and confirmed that inclusions of the bilateral filter, CLAHE cGAN, EfficientNetB5, and the CBAM attention mechanism significantly improved accuracy from 97.04% to 98.81%. Based on the findings, the study concludes that thermal imaging is a highly reliable modality for emotion recognition, especially in environments where visible light cameras fail. By successfully integrating cGAN-generated synthetic data and attention mechanisms (CBAM), the proposed model effectively overcomes the traditional limitations of small datasets and low-contrast imagery. Despite using different datasets, our proposed approach demonstrates promising performance compared to related studies.

All the Declarations and Statements

Author Contributions Statement

T.A: Handled data acquisition, dataset preprocessing, Model Training, Validation, Visualization, prepared performance charts, and benchmarked performance against existing methods. S.H. Conceptualization, Methodology, and Supervision: Proposed research ideas, constructed the overall framework, supervised the project, ensured the statistical robustness of the evaluation, and helped coordinate project milestones and deadlines. K.A: Writing– Drafted the initial manuscript, contributed to the literature survey, documented the technical background of the study, reviewed and edited the manuscript, ensured clarity and coherence, Formal Analysis.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The data supporting the results of this study are included in the article. Additionally, the Data for this article are available at “<https://data.mendeley.com/datasets/8885sc9p4z/1>”, accessed on “November 12, 2024”.

Ethical Declarations

N/A

Acknowledgements

We wholeheartedly appreciate the professional analysis provided by the experts and their useful suggestions, which have helped to enhance the quality of the experiment and the validity of its findings.

Declaration of Generative AI in Scholarly Writing

Language polishing and grammar were done with the help of AI. There was no use of a generative AI to create content, generate ideas, analyze data, generate figures, or write scientifically.

Abbreviations

This manuscript uses the following abbreviations:

cGAN-Conditional Generative Adversarial Network
CLAHE-Contrast Limited Adaptive Histogram Equalization
CBAM-Convolutional Block Attention Module
FER-Facial Expression Recognition
FID-Fréchet Inception Distance
GAN-Generative Adversarial Network
MBConv-Mobile Inverted Bottleneck Convolution
ROC AUC-Receiver Operating Characteristic Area Under the Curve
TP – True Positive
FP – False Positive
TN – True Negative
FN – False Negative
F1-score – Harmonic mean of precision and recall

Appendix A/B/C..., with appendix title

None.

References

- [1] A. H. Krishna, A. B. Sri, K. Y. V. S. Priyanka, S. Taran, and V. Bajaj, "Emotion classification using EEG signals based on tunable-Q wavelet transform," *IET Science, Measurement and Technology*, vol. 13, Iss. 3, pp. 375–380, 2019, doi: 10.1049/iet-smt.2018.5237
- [2] E. S. Agung, A. P. Rifai, and T. Wijayanto, "Image-based facial emotion recognition using a convolutional neural network on Emotion Recognition Dataset," *Scientific Reports*, vol. 14, 14429 (2024), <https://doi.org/10.1038/s41598-024-65276-x>
- [3] V. Jacintha, J. Simon, S. Tamilarasu, R. Thamizhmani, K. T. Yogesh, and J. Nagarajan, "A review on facial emotion recognition techniques," in *2019 International Conference on Communication and Signal Processing (ICCSP)*, 2019, pp. 517–521, DOI: 10.1109/ICCSP.2019.8698067.
- [4] E. G. Dada, D. O. Oyewola, S. B. Joseph, O. Emebo, and O. O. Oluwagbemi, "Facial emotion recognition and classification using the convolutional neural network-10 (CNN-10)," *Applied Computational Intelligence and Soft Computing*, vol. 2023, PP.19, 2023, DOI: <https://doi.org/10.1155/2023/2457898>
- [5] A. Seal, S. Ganguly, D. Bhattacharjee, M. Nasipuri, and D.K. Basu, "Thermal Human Face Recognition Based on Haar Wavelet Transform and Series Matching Technique," In: Swamy, P., Guru, D. (eds) *Multimedia Processing, Communication and Computing Applications. Lecture Notes in Electrical Engineering*, vol 213. Springer, New Delhi. https://doi.org/10.1007/978-81-322-1143-3_13
- [6] A. Bhattacharyya, S. Chatterjee, S. Sen, A. Sinitca, D. Kaplun, and R. Sarkar, "A deep learning model for classifying human facial expressions from infrared thermal images," *Scientific Reports*, vol. 11, 20696 (2021), <https://doi.org/10.1038/s41598-021-99998-z>
- [7] Chaitanya, S. Sarath, Malavika, Prasanna, and Karthik, "Human Emotions Recognition from Thermal Images using Yolo Algorithm," *2020 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, 2020, pp. 1139–1142, doi: 10.1109/ICCSP48568.2020.9182148.
- [8] A.K. Prabhakaran, J.J. Nair, S. Sarath, "Thermal facial expression recognition using modified ResNet152," *Advances in Computing and Network Communications, Lecture Notes in Electrical Engineering* 736, 2021. https://doi.org/10.1007/978-981-33-6987-0_32
- [9] B. Assiri and M. A. Hossain, "Face emotion recognition based on infrared thermal imagery by applying machine learning and parallelism," *Mathematical Biosciences and Engineering*, vol. 20, pp. 913–929, 2023, doi: 10.3934/mbe.2023042.
- [10] Yomna M. Elbarawy, Neveen I. Ghali, Rania Salah El-Sayed, "Facial Expressions Recognition in Thermal Images based on Deep Learning Techniques", *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, Vol.11, No.10, pp. 1-7, 2019. DOI: 10.5815/ijigsp.2019.10.01
- [11] S. B. R. Prasad and B. S. Chandana, "Human face emotions recognition from thermal images using DenseNet," *International Journal of Electrical and Computer Engineering Systems*, vol. 14, pp. 155–167, 2023, DOI: <https://doi.org/10.32985/ijeces.14.2.5>.
- [12] C.K. Kyal, H. Poddar, H., M. Reza. "Human Emotion Recognition from Spontaneous Thermal Image Sequence Using GPU Accelerated Emotion Landmark Localization and Parallel Deep Emotion Net," *International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing*, vol 1165, 2020. Springer, Singapore. https://doi.org/10.1007/978-981-15-5113-0_78
- [13] Abuhussein, Mohamed Fawzi; Darwish, Ashraf; Hassanien, Aboul Ella. "Comprehensive Facial Thermal Dataset", *Mendeley Data*, V1, 2024, doi: 10.17632/8885sc9p4z.1
- [14] P. A. Flach, "The geometry of ROC space: Understanding machine learning metrics through ROC isometrics," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, Washington, DC, USA, 2003, pp. 194–201.
- [15] H. Jiang, X. Chen, S. Chen, L. Ye, J. Hu and K. Yang, "Beyond the Visible: Deep Learning-Powered Thermal Face Recognition," in *IEEE Internet of Things Journal*, vol. 13, no. 7, pp. 14867–14881, 1 April, 2026, doi: 10.1109/JIOT.2026.3652708.

Authors' Profiles



Tahir Aman received his B.Sc. degree in Information Systems from Haramaya University, Ethiopia, in 2017, and his M.Sc. in Computer Science and Engineering from Adama Science and Technology University (ASTU), Ethiopia, in 2025. He is currently a Lecturer and researcher in the Information Systems Department, Werabe University, Ethiopia. His research interest areas are Computer Vision, face recognition, Information processing and management, and deep learning.



Sintayehu Hirpassa, PhD, received his B.Sc. degree in Information Science from Jimma University, Ethiopia, in 2009, his M.Sc. in Information Science from Addis Ababa University, Ethiopia, in 2013, and his Ph.D. in Computer Science from Punjabi University, India, in 2017. He is currently a Lecturer and researcher in the Computer Science and Engineering Department at Adama Science and Technology University (ASTU), Ethiopia. His research interest areas are Natural Language Processing, Information Extraction, Machine Learning, Text Classification, data mining, and computer vision.



Kibreab Adane, PhD, received his B.Sc. degree in Information Science from Jimma University in 2008, his M.Sc. in Information Science from Addis Ababa University in 2013, and his Ph.D. in Computing and Information Technology, AI in Cybersecurity from Arba Minch University in 2024. He is currently a Lecturer and researcher in the Computing and Software Engineering Faculty at Arba Minch University. He is a member of Technical Peer Review Committees in several reputable journals. His research interest areas are AI, Cybersecurity, Machine Learning, Deep Learning, Text Classification, Data Mining, Computer Vision, and wired and wireless networks

How to cite this paper

Tahir Aman, Sintayehu Hirpassa, Kibreab Adane, "Thermal Emotion Recognition with and without Synthetic Facial Images using Attention-Based EfficientNet5", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.4, pp.211-227, 2026. DOI:10.5815/ijem.2026.04.15