Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 10, No. 2, April, 2026

Detecting Genuine Versus Fake Emotions: A Dual-Task Deep Learning Approach Using Facial Expression Analysis

Sarah Tasnim Diya [1], Most. Jannatul Ferdos [1], Md. Mizanur Rahman [1*], Yadab Sutradhar [2], Zahura Zaman [3*], Suman Ahmmed [4], Ohidujjaman [4]

[1] *Faculty of Science and Information Technology, Daffodil International University, Dhaka 1216, Bangladesh.*

[2] *Department of Computer Science, Maharishi International University, Fairfield, Iowa, United States.*

[3] *Department of Computing, Boise State University, Idaho, United States.*

[4] *Department of Computer Science and Engineering, United International University, Dhaka 1212, Bangladesh.*

Abstract

Facial expression recognition (FER) is a relevant field of study with applications in human-computer interaction, healthcare, and security. Although recent approaches demonstrate excellent outcomes on the recognition of basic emotions, the authenticity of expressions (genuine versus fake) remains unexplored. In this work, we propose a dual-task deep learning framework based on EfficientNet-B0, enhanced with a lightweight squeeze-and-excitation (SE) attention mechanism, to collaboratively work on multiclass emotion recognition (seven categories: angry, disgust, fear, happy, neutral, sad and surprise) and authenticity classification (genuine vs fake). The architecture leverages a shared backbone for representing feature, followed by task-dedicated branches trained using categorical cross-entropy and focal loss, respectively. To overcome the lack of publicly available benchmarks incorporating authenticity labels, we designed a curated dataset annotated with both emotional categories and authenticity information. Experimental evaluation demonstrates that the proposed dual-task model with the SE attention mechanism achieves 98.5% accuracy for emotion recognition and 92.2% accuracy for authenticity prediction, emphasizing both the effectiveness of the framework and the inherent challenges of authenticity detection. Moreover, we present a deployable real-time system demonstrating the feasibility of integrating authenticity-aware FER into practical applications such as e-learning analytics, security surveillance, and affective computing.

Keywords:

Affective Computing;
Authenticity Detection;
Dual-Task Learning;
EfficientNet; Facial Emotion;
Emotion Recognition;
Attention Mechanisms;
Deep Learning.

Article History:

Received:	29	December	2025
Revised:	03	March	2026
Accepted:	11	March	2026
Published:	01	April	2026

1- Introduction

Facial emotion recognition (FER) has been a keystone in affective computing for decades with applications such as healthcare, education, surveillance, and human-computer interaction [1]. Conventional FER work mainly targets at recognizing the discrete emotions such as happiness, anger or sadness, achieving high classification accuracy on well-controlled data [2, 3]. But a significant challenge remains unaddressed: distinguishing if an expression is real or fake. Such an authenticity dimension is important in real-world applications, such as security screening, lie detection and trust evaluation in human–AI communication.

While much of the existing FER methods have been developed focusing on emotion recognition alone, independent of authenticity, thereby overlooking the fact that two visually similar expressions of the same emotion may differ

* **CONTACT**: mizanurrahman.cse@diu.edu.bd; zahurazaman@u.boisestate.edu

DOI: https://doi.org/10.28991/ESJ-2026-010-02-018

© 2026 by the authors. Licensee ESJ, Italy. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC-BY) license (https://creativecommons.org/licenses/by/4.0/).

fundamentally in whether they are genuine or posed [4]. Even if the authenticity is taken into account, it is often conceived of as a separate binary problem, which is independent of emotion detecting [5]. This differentiation disregards the inherent association of the emotional category with actual and performed versions. As an example, minor micro-expressions that appear in actual anger are very different with exaggerated expressions that appear in simulated anger. This kind of interrelationship may be learned by employing a multi-task environment, which learns both shared facial patterns and task-related information at the same time.

In order to fill this gap, we develop a dual-task attention-enhanced model that recognizes emotions and detects authenticity at the same time. According to EfficientNet-B0, the model combines with lightweight attention layers where attention is paid to discriminative facial parts, and the multi-task structure allows sharing in learning between the two outputs. Experimental validation is conducted using a newly collected dataset of genuine and acted expressions over seven basic emotions.

The main contributions of this paper can be summarized as:

- A novel multitask architecture that simultaneously performs emotion recognition and authenticity detection in a single framework.
- Introduction of an attention mechanism to enhance the differentiability of subtle emotional micro-patterns.
- A curated primary dataset with authenticity annotations, to enable a more fine-grained evaluation.
- Integration of Explainable AI methods (LIME and Grad-CAM) to ensure the model transparency and comprehension of the decision made by the model.
- An actual-world application pipeline demonstrating the feasibility with which this approach could be scaled.

The rest of this paper is organized as follows: We introduce related work in Section 2, the data and preprocessing are described in Section 3, the proposed model architecture is detailed in Section 4, experimental results and analysis are presented in Section 5, explainable AI are discussed starting from section 6; then we conclude with conclusion highlighting the future work in Section 7.

2- Related Works

Emotion recognition and more specifically, differentiation of genuine from posed facial expressions has been a topic of interest in the domains of computer vision, affective computing and psychology for some time. The work on approaches ranges from psychological based, through cutting-edge deep learning systems, multimodal approaches and explainable AI. This section presents a brief review of the advancements that have been made by prior works in FER and genuine detection, in addition to uncovering a few gaps and justifying our work.

2-1- Formative Years of Emotion Authenticity

The cognitive basis for separating real and feigned emotions was established in Ekman's ground breaking research on universal facial expressions alongside the Facial Action Coding System (FACS) [6]. Building on this, Jia et al. [7] also presented a systematic review of the datasets and methods with a focus on micro-expressions for authentic emotion. This work and other early works were the groundwork for the development of computational methods later used to detect authenticity.

2-2- CNN-Based Emotion Recognition

Recent developments in deep learning caused the convolutional neural networks (CNNs) to dominate FER. Bhagat et al. [8] applied CNNs to FER2013, achieving a training accuracy of 82.56% and a very low verification accuracy of 65.68% because of class bias, overlapping expressions, and noise in images. Similarly, Ballesteros et al. [9] showed that image quality and context information significantly affect the recognition performance.

Hybrid solutions, such as integration of CNN and traditional features, have also been considered. Mathur & Gupta [10] incorporated Local Binary Patterns (LBP) with the Gabor filters to increase the performance on low-resolution images. Anand & Babu [11] performed optimization of EfficientNetB0 using Red Fox Optimizer, which resulted in superior performance on FER2013 and EMOTIC. Manimohan et al. [12] utilized ResNet50 with Haar cascades for real-time recognition on webcams and also obtained competitive results. In another work, Khuntia & Kale [13] demonstrated that CRNN architecture enhances the temporal learning with an accuracy of 79.72% over FER2013+.

2-3- Model Architectures and Lightweight Designs Advancements

Towards higher performance and better efficiency, some advanced structures have been introduced as well. Islam et al. [14] used Neural Architecture Search (NASNet) combined with reinforcement learning, achieving more than 98% accuracy in FER2013 and CK+. Haider et al. [15], combining CNNs, SVM classifiers, and triplet loss, presented the best results on JAFFE and MMI datasets, but less generalization on AFFECTNET.

Light-weight and real-time models have also emerged. Singh [16] presented a land-mark based approach with 84.27% accuracy, which is sensitive to generic head rotation. Similarly, Hakim et al. [17] designed CCR based OpenCV pipeline, achieving an accuracy of ~85% on real-world surveillance and retail.

2-4- Video-Based and Temporal Approaches

Video FER is able to capture the temporal characteristics of expressions. Ashraf et al. [18] employed a CNN on the ADFES-BIV collection and achieved an accuracy of 99.38%, but reported overfitting as an issue. Pruthviraja et al. [19] proposed FER2013+ to expand DCNNs and demonstrated the ethical issues of data collection, achieving an accuracy of 81.33%. Explainability has also become important. Cardaioli et al. [20] utilized the SASE-FE dataset and explainable AI methods to discover facial movements of muscles, which are important for authentication. Meanwhile, Miolla et al. [21] proposed PEDFE, a dataset for both dynamic genuine and posed expressions annotated by experts, which facilitated making more realistic databases.

2-5- Authentic and Fake Emotion Detection

Recent studies have focused on authenticity detection, separating authentic from acted expressions. Annadurai et al. [22] introduced an Enhanced Boosted SVM with 98.08% results on SASE-FE as well as FED datasets. Sunil et al. [23] exploited temporal cues using modified CNNs and achieved 96% accuracy over the ChaLearn and Fake Smile Master datasets. Fake emotion detection has also been advanced based on multimodal approaches. Arslan et al. [24] used ECG & GSR in VR environment, where + 97.78% accuracy is achieved. Jia et al. [25] projected emotion into Valence-Arousal-Dominance (VAD) space with the help of multiple modalities for cross-validation. Govea et al. [26] employed CNN-RNN models with reinforcement learning to personalize learning environments and achieved an accuracy of 88% over biometric, face, and speech data.

2-6- Domain-Specific Applications

Detecting fake emotions is also applied in the field of education and health care. Evangeline & Parkavi [27] used InceptionV3 for online learning datasets and achieved 96.26% accuracy. Rathod et al. [28] combined social media, speech, and facial cues for remote MH monitoring and achieved accuracies of 92% (facial) and 87.96% (speech). Chethan & Vinay [29] developed a robust CNN-RNN model against occlusions and noise, while Barnwal & Barik [30] applied Haar-based classifiers for group emotion analysis.

Additional hybrid contributions have been reported, such as Zhang [31], fused ResNet and MobileNet (80% accuracy, but not good in class-wise performance), Singh et al. [32] have used real-time FER in behavioral learning analysis, and Ton-that & Cao [33] employed Combined Gray LBP (CGLBP) method using SVM, which has reached up to 99% over JAFFE and MUG datasets.

Table 1 provides a summary of recent and prominent past studies on emotion and authenticity recognition, starting with Jia et al. [7], which was the first article to propose the fundamental taxonomy of genuine-posed datasets and then the recent CNN, hybrid, and multimodal frameworks.

Table 1. Summary of related works on facial emotion and authenticity recognition

Ref.	Year	Method / Model	Dataset Used	Key Contribution	Acc. / Result
Jia et al. [7]	2021	Review of genuine vs. posed emotion databases	Multiple (e.g., CK+, MMI)	Surveyed databases and methods for genuine vs. posed emotion detection	-
Bhagat et al. [8]	2024	CNN	FER2013	Identified Class bias and low generalization in FER models	65.68% (val)
Mathur & Gupta [10]	2024	CNN + LBP + Gabor	Custom image set	Hybrid handcrafted–deep features improved low-res emotion detection	90%
Anand & Babu [11]	2024	EfficientNet-B0 (optimized)	FER2013, EMOTIC	Meta-heuristic optimization improved model generalization	96%
Islam et al. [14]	2023	NASNet (RL-based)	FER2013, CK+	Auto-designed CNN achieved state-of-the-art FER accuracy	98%
Haider et al. [15]	2023	CNN + SVM + Triplet Loss	JAFFE, MMI	Metric learning improved intra-class separability	95%
Ashraf et al. [18]	2023	Video CNN	ADFES-BIV	Modeled temporal cues in dynamic expressions	99.38%
Cardaioli et al. [20]	2022	Explainable CNN	SASE-FE	Applied XAI to interpret micro-expressions for genuineness	96%
Miolla et al. [21]	2022	PEDFE dataset	PEDFE	Introduced dynamic benchmark labeled for genuine vs. posed	-
Annadurai et al. [22]	2022	Boosted SVM	SASE-FE, FED	Enhanced SVM classifier for real/fake emotion detection	98.08%
Sunil et al. [23]	2023	Modified CNN	ChaLearn, Fake Smile Master	Classified real vs. fake emotions using temporal cues	96%
Arslan et al. [24]	2024	Multimodal (ECG + GSR)	Custom VR dataset	Biosignal-based multimodal authenticity detection	97.78%
Jia et al. [25]	2024	Multimodal VAD	Multiple cross-domain	Modeled emotions in valence–arousal–dominance space	-

These recent works represent a milestone in distinguishing true and false emotion credibility based on deep learning, hybrid network and multimodal learning. There is, however, a large performance gap in real emotion recognition, partly because natural expressions are secretly known to be variant among individuals, and that they are culturally conditioned. The current detection systems, despite the improvements in accuracy, robustness, and adaptability, still possess long-standing challenges, particularly under the conditions of unrealistic and unbalanced datasets with an authenticity label, the difficulties in more subtle micro-expression reading, and the inability to generalize across domains of current models. To fill these gaps, more sustainable work is necessary to better diversify the data and work on the interpretability of the models and domain-adaptation strategies, which are to be used to create more scalable and reliable emotion-detecting systems.

3- Dataset Development

The development of a reliable dataset was crucial to the dual-task challenge of recognizing both primary emotions and their authenticity. Since there was no publicly available dataset that sufficiently captured the interrelation of emotion categories and fake vs. real facial expressions, we introduced a customized dataset for the purpose of this study.

The dataset includes seven basic emotional categories (Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise), each of which was divided into 2 expressional classes: Genuine and Fake. This architecture allowed to learn both categorical emotion classification and authenticity analysis concurrently.

3-1- Data Collection

Data were collected in a controlled setting and across diverse participants. The primary dataset consisted of 46 subjects (28 male, 18 female), including 7 professional actors to provide more plausible, posed expressions. The sample consisted of individuals from a wide range of ages, genders, colors, and geographic locations. Genuine emotions were evoked by natural inductions (e.g., engaging videos or personal recollections), and fake expressions were elicited with directed acting tasks.

Because it is difficult to obtain spontaneous emotion data, we made an attempt to create natural settings for genuine classes. Subjects were instructed to feel emotions as spontaneously as possible, while posed categories were exaggerated in order to reach the level of acted emotion.

All data collection and annotation procedures were performed under the supervision of a licensed psychologist for compliance with ethical regulations, to reduce possible discomfort among participants, and to confirm the genuineness of emotional responses. Written consent was taken from all the individuals before collecting the data.

3-2- Data Labeling and Annotation Protocol

Authenticity labels were assigned under the supervision of a licensed psychologist following standardized facial expression criteria. To ensure labeling consistency, a subset of samples was independently re-evaluated at different intervals, yielding high agreement consistency. While formal multi-annotator Cohen's Kappa was not computed due to expert-guided single-annotator supervision, this protocol ensured reliable and psychologically grounded authenticity annotations. Micro-expressions and subtle cues were given specific consideration by the following guidelines:

- Eye involvement: The muscles around the eyes are involved in genuine emotions but may not be involved in fake emotions.
- Symmetry: Natural emotions are a bit asymmetrical; pretended emotions are totally symmetrical.
- Consistent Body-language: Authentic expressions coincide with form, posture and movement of lips
- Brief micro-expressions: Genuine emotions are usually a fraction of a second long even repressed.

The detailed criteria used for each emotion class are summarized in Table 2.

Table 2. Labeling guide for authenticity across the seven basic emotion classes

Emotion	Genuine Indicators	Fake Indicators
Happy	Eye wrinkles, natural cheek lift, symmetrical smile	Mouth-only lift, asymmetrical or abrupt smile
Sad	Raised inner brows, drooping eyelids, slow onset	No drooping, mismatched eyes–mouth, sudden onset
Angry	Eyebrow tension, tight lips, clenched jaw	Exaggerated brows, missing lip/jaw tension
Fear	Raised eye brows, eyelids pressed tight together, brief open mouth	Incorrect brows, mismatched mouth, prolonged expression
Surprise	High brows, brief jaw drop	Excessive jaw drop, long-held, unsynced brows
Disgust	Curled nostrils, uplifted lip and squeezed eye lids	Mouth curl only, missing wrinkles, exaggerated curl
Neutral	Subtle asymmetry, slight eye narrowing	Forced symmetry, lips raised unnaturally long

Cues are based on psychological micro-expression research and tested with the guidance of a licensed psychologist.

3-3- Dataset Statistics

Overall, the dataset included seven basic emotions (Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise), each with two subsets (i.e., Genuine and Fake). The dataset structure consisted of a hierarchical folder arrangement, with higher-level emotion classes and lower-level authenticity subfolders. In the initial step, a total of 2,224 raw samples were collected across the seven emotion categories (real or fake). After refinement, a subset of 1,657 samples was selected for modeling, taking into account class imbalance, noise, and ambiguous cases. To ensure a fair performance comparison, the dataset was split into a training set of 1,325 samples (80%) and a test set of 332 samples (20%). Table 3 presents the distribution of the primary dataset across different categories.

Table 3. Final distribution of samples across emotion and authenticity categories

Emotion	Fake	Genuine	Total
Angry	99	64	163
Disgust	152	115	267
Fear	134	104	238
Happy	152	184	336
Neutral	141	169	310
Sad	76	125	201
Surprise	101	41	142
Total	**855**	**802**	**1657**

A total of 1657 samples were selected from a pool of 2224 collected samples. The total number of images was 1657, with 1325 for training and 332 for testing.

3-4- Preprocessing and Augmentation

For training, images were resized to 224 × 224 pixels and pre-processed accordingly with input preprocessing functions consistent with the EfficientNet architecture family. Data augmentation (i.e., random horizontal flips, minor rotations, contrast and brightness changes) was used to generalize the model and reduce overfitting. These operations also helped to maintain natural features in facial expressions, and improved generalization.

Despite SMOTE is conventionally applied to tabular data, in this work it was applied on learned feature embeddings extracted from the backbone network rather than on raw image pixels. This approach has been adopted in prior deep learning studies where embeddings form a structured feature space. Alternative strategies such as focal loss alone were also considered; however, empirical validation showed that combining focal loss with embedding-level SMOTE improved minority-class stability without introducing visual artifacts. The dataset preparation workflow, illustrated in Figure 1, outlines the process from initial collection and refinement to final augmentation.

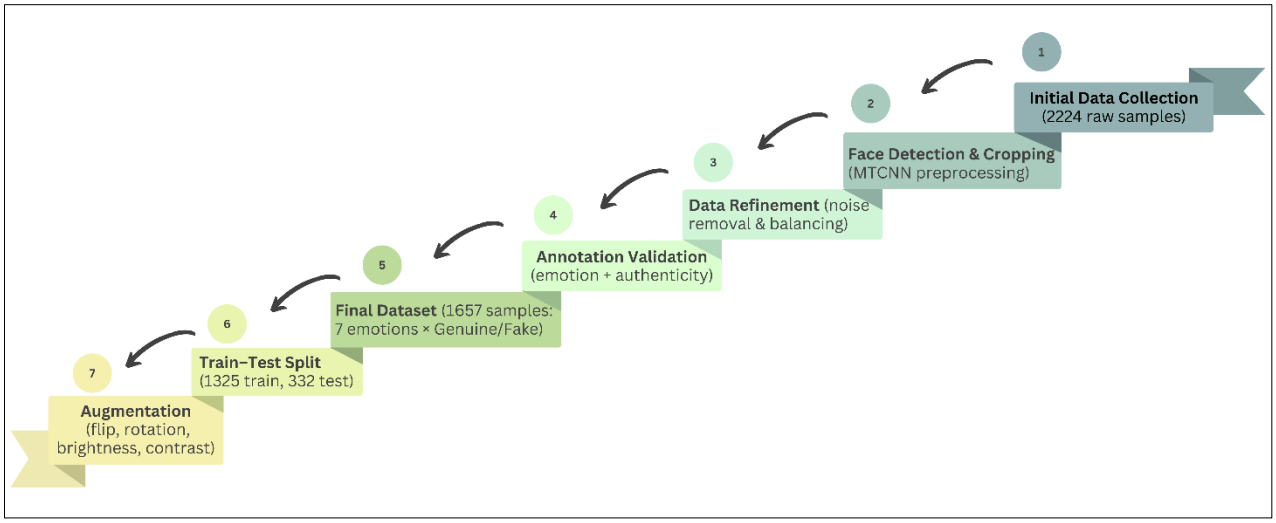


Figure 1. Dataset preparation and augmentation workflow

Figure 2 depicts representative samples of both authentic and counterfeit facial expressions of seven emotion categories, (Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise demonstrating variations in facial muscle activation and movement dynamics observed to discriminate between authenticity labels during annotation.

Figure 2. Sample genuine and fake emotion examples across seven categories

4- Research Methodology

The proposed model aims to combine facial emotion recognition with authenticity detection into a single deep learning framework. Differing from traditional single-task model, our framework benefits from a joint feature backbone and task-specific branches, which offer efficient learning of both categorical emotions and authenticity cues.

4-1- Model Architecture

We utilized EfficientNet-B0 pretrained on ImageNet as the backbone encoder, due to its efficiency–accuracy trade-off. The normalized 224×224×3 input images were passed through the encoder to generate condensed embeddings using global average pooling. These embeddings were then fed through a couple of task-specific branches.

The emotion recognition branch had a fully connected layer projecting to seven output units with SoftMax, generating probabilities over the emotion categories (angry, disgust, fear, happy, neutral, sad, surprise).

The authenticity branch, on the other hand, contained a fully connected layer that mapped to a single output neuron with sigmoid activation, to differentiate real and fake expressions. Dropout layers were further used in the task-specific branches to mitigate overfitting during training. Moreover, a lightweight attention module named squeeze-and-excitation (SE) was embedded into the backbone to improve discriminative capacity by focusing on important regions of faces such as eyes, brows, and mouth. Additionally, stochastic depth regularization was incorporated within the backbone to improve generalization. The squeeze-and-excitation (SE) attention modules are embedded within the MBConv blocks of the EfficientNet-B0 backbone, following depthwise convolution and before residual connections. This placement enables channel-wise recalibration of intermediate feature maps, emphasizing discriminative facial regions relevant to both emotion and authenticity tasks.

The proposed architecture assumes a meaningful interdependence between facial emotion categories and expression authenticity, which motivates joint feature learning through a shared backbone. The interdependence between emotion category and authenticity was empirically examined through ablation experiments. When the authenticity branch was removed, emotion classification performance declined, particularly for subtle expressions such as fear and sadness. Conversely, authenticity prediction accuracy deteriorated when emotion supervision was excluded. These findings indicate that shared feature representations capture complementary cues across both tasks, supporting the assumption that emotion category and authenticity are jointly informative rather than independent learning objectives.

The overall workflow of the proposed dual-task model, including shared feature extraction and the two parallel output branches, is illustrated in Figure 3, which outlines the architecture's hierarchical flow from preprocessing to final classification outputs.

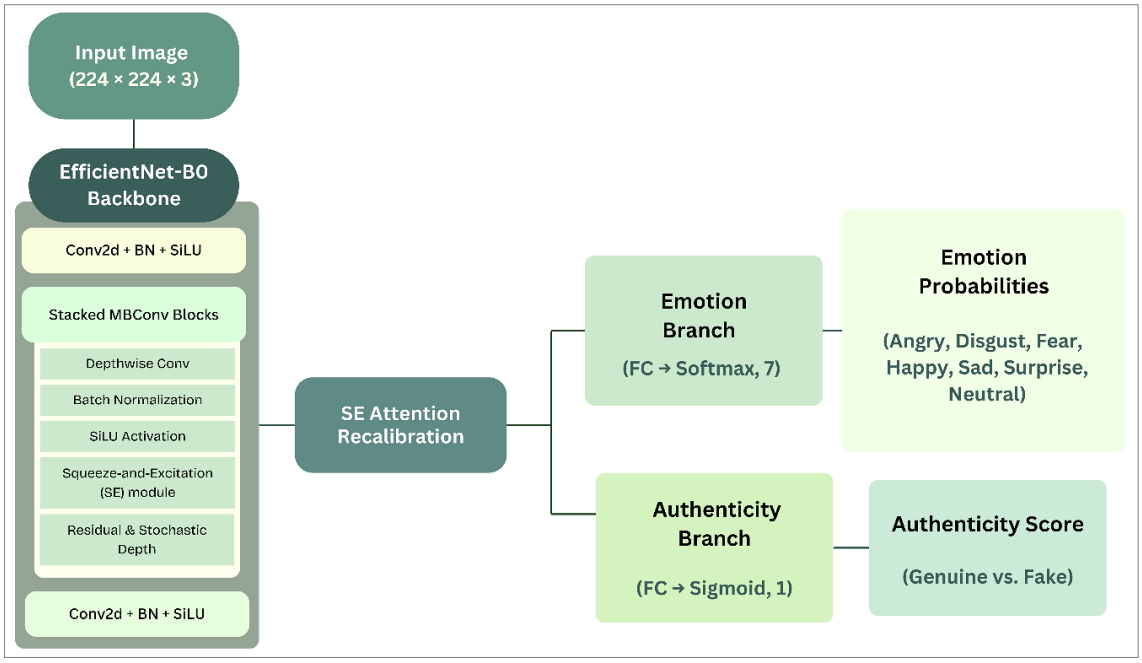


Figure 3. Proposed dual-task architecture using EfficientNet-B0 with SE attention

4-2- Training Procedure

To mitigate the distributional imbalance in authentic labels, SMOTE-based oversampling was used. Training was done using Augmentation methods such as horizontal flips, random rotations or brightness changes to enhance robustness and avoid overfitting. The training loss function integrated the following two components, categorical cross-entropy for emotion recognition and BCE for authenticity classification.

$$L_{total} = \alpha \cdot L_{emotion} + \beta \cdot L_{authenticity} \quad (1)$$

where we fixed the weights as α = 0.4 and β = 0.6 to emphasize the authenticity task. We optimized our model using the Adam optimizer with learning rate of 1 $\times 10^{-4}$ and ReduceLROnPlateau scheduler. Training was done for maximal 100 epochs with early stopping (patience = 10).

The main hyperparameters and training setup are presented in Table 4.

Table 4. Training hyperparameters and model configuration

Parameter	Value
Backbone	EfficientNet-B0
Input Size	224 × 224 × 3
Batch Size	32
Optimizer	Adam
Initial LR	1e-4
LR Scheduler	ReduceLROnPlateau
Epochs	100 (early stopping=10)
Loss Functions	Cross-Entropy, BCE
Loss Weights	α = 0.4, β = 0.6

Here, α and β denote weighting factors applied to balance the emotion recognition and authenticity classification losses, respectively. Multiple weighting combinations were initially evaluated during preliminary experiments. Balanced settings (α = β = 0.5) resulted in reduced authenticity sensitivity, while higher emotion weighting led to degraded genuineness detection. The selected values (α = 0.4, β = 0.6) provided the best trade-off between stable emotion classification and improved authenticity discrimination, and performance was observed to be relatively stable within a ±0.1 range.

The entire experimental procedure of the suggested dual-task framework, as well as data preparation, preprocessing, model building, training-testing, and evaluation, is outlined in Figure 4, which shows the sequential pipeline starting with the data acquisition and culminating in the final emotion and authenticity prediction.

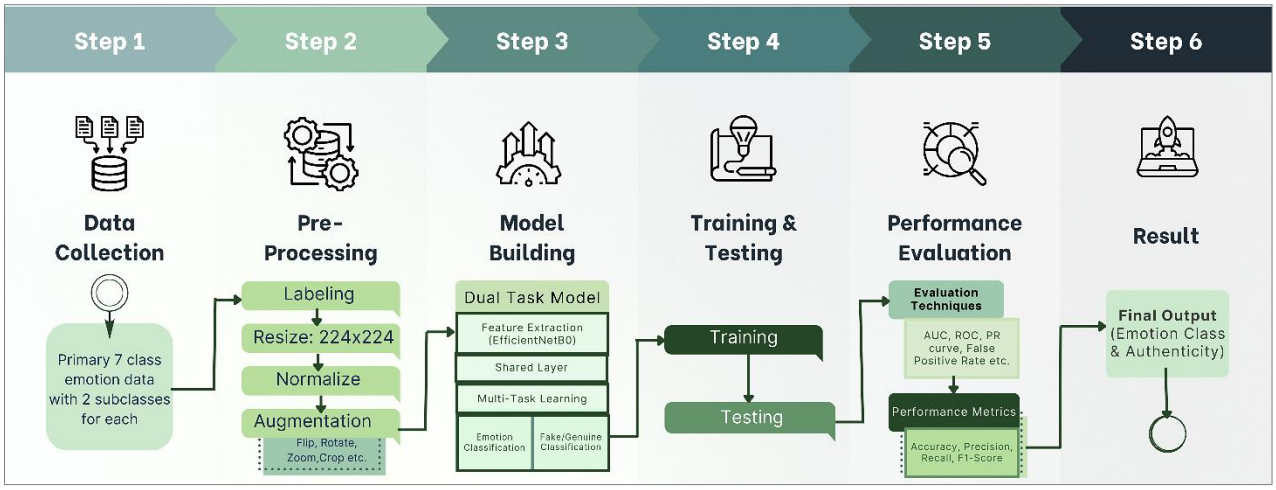


Figure 4. Methodology pipeline of the proposed dual-task model

4-3- Deployment Pipeline

The emotion and genuineness recognition dual-task model employs an EfficientNet-B0 backbone with squeeze-and-excitation (SE) attention layer. It has dual-outputs: (i) a SoftMax head for seven basic emotions (Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral), and (ii) a sigmoid activation head for the authenticity label (authentic vs. fake). The best checkpoint (.pth) was preserved for inference.

The backend, built using FastAPI, has face detection based on the MediaPipe framework by Google. The detected faces are resized to 224×224 pixels, and normalized before being fed to the network as a tensor. The predicted probabilities of emotions and authenticity are produced by the model, with an authenticity score thresholded at 0.5. Output bounding boxes and predictions are streamed to the frontend using WebSockets for real time on demand frame analysis.

The frontend written in HTML5 and JavaScript receives live video and enables the submission of frames manually to be analyzed. The interface has the 3-step sequence, rather than the streaming, which involves Starting the camera, connecting with the server, and Test Frame. Visualization of bounding boxes with predicted emotions and authenticity labels is superimposed and frames can be saved to be examined qualitatively.

The system is deployed directly onto Render without being containerized. It is based on FastAPI and Uvicorn servers and CORS middleware to provide a secure means of communication. The entire codebase is broken down into individual modules (e.g. model.py, app.py, index.html) with a requirements.txt containing the list of dependencies needed to ensure seamless deployment across cloud environments.

In Figure 5, a real-time deployment pipeline is presented, illustrating the sequential flow of live frame capture, backend processing, model inference, and frontend visualization.

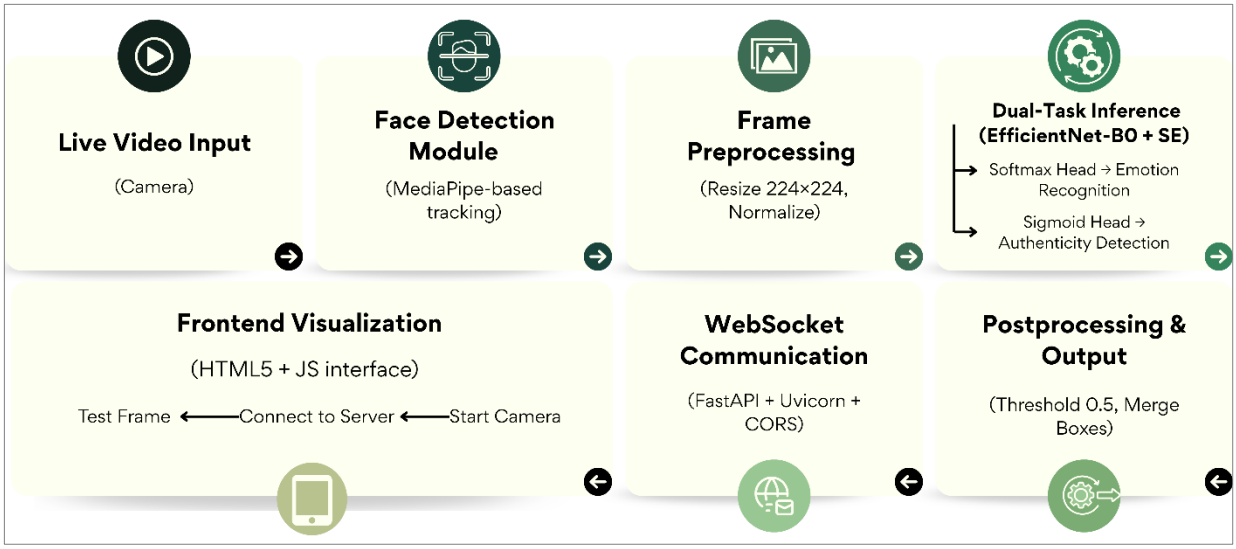


Figure 5. Deployment pipeline for real-time inference

4-4- Comparative Model Architectures

In order to guarantee reliability and fairness of evaluation, a series of baseline and hybrid models were created with the same preprocessing, training, and testing conditions. The models were initialized with 224 × 224 × 3 normalized face inputs and trained with Adam optimizer (learning rate = 1×10^{-4}) in its joint loss configuration, which matched the proposed framework. The following architectures were taken to have comparative evaluation:

4-4-1- CNN (Baseline)

A lightweight convolutional neural network that is built as a benchmark of dual-task learning. The model is made up of three convolutional blocks, and each block comprises a convolutional layer of 3 × 3 with ReLU activation and a max-pooling layer to do spatial downsampling. The resulting feature maps are flattened and fed through a single fully connected 128-neuron fully connected layer with a dropout rate of 0.5 to regularize them. Two parallel output heads are then implemented, namely, a seven-unit SoftMax layer that identifies emotion and a one-unit sigmoid layer that identifies authenticity. The model is trained at once on both tasks on categorical cross-entropy and binary cross-entropy loss, with a total number of trainable parameters approximately equal to 3.3 million.

The schematic structure of this baseline model is illustrated in Figure 6, highlighting its simple convolutional design and dual-output configuration.

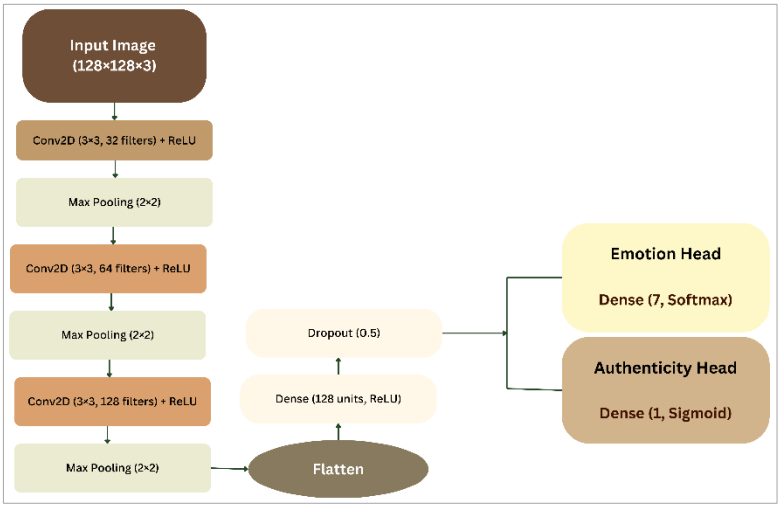


Figure 6. Baseline CNN model architecture for dual-task classification

4-4-2- EfficientNet-B0 + MobileNetV2 (Hybrid Branch)

A two-stream hybrid network with EfficientNet-B0 to identify emotions and MobileNetV2 to detect authenticity. Normalized 224 × 224 × 3 inputs were processed by each branch to generate global pooled embeddings, which were then fed through dense layers (256 units in the case of emotion, 128-64 units in the case of authenticity) with ReLU activations and dropout regularization. The model was optimized with a combination of the categorical cross-entropy and binary focal loss ($\gamma = 2$) with loss weights of 1.0 and 1.5, respectively. Around 6.8 million parameters were applied, of which 4.33 million were trainable. This design evaluated the effectiveness of separate feature encoders for each subtask. The architecture is depicted in Figure 7, which outlines the parallel feature encoding and distinct dense layers for each task.

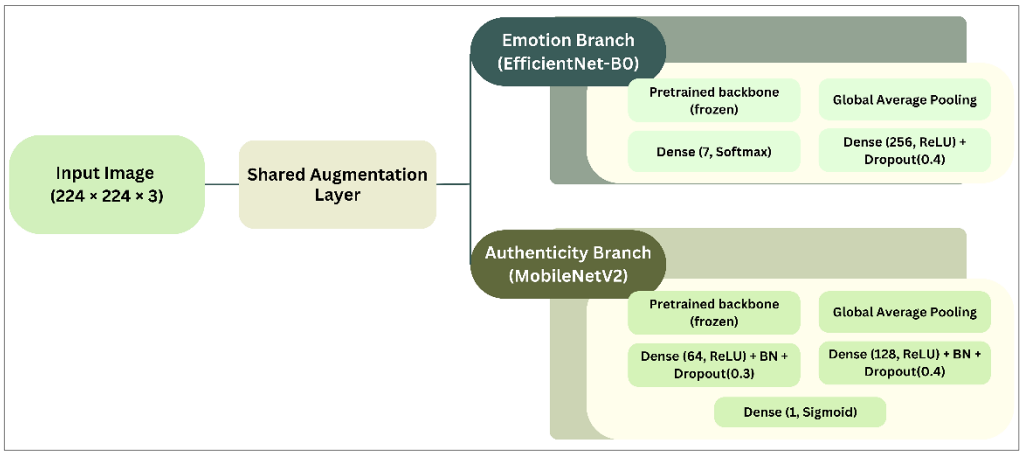


Figure 7. Hybrid dual-branch EfficientNet-B0 + MobileNetV2 architecture

4-4-3- EfficientNet-B3 (Feature Embedding + Dense Classifier)

A configuration with featured embedding, where a pretrained EfficientNet-B3 backbone (frozen by the time of training) was taken as a high-level feature extractor. Each image was converted to 224 x 224 x 3 and normalized and fed through the backbone to generate 1536-dimensional global pooled embeddings. This was then fed to a dense layer of 256 neurons, which had ReLU activation and a dropout rate of 0.3 to be regularized. Two output heads were connected independently: a seven-unit SoftMax classifier used to identify emotions and a one-unit sigmoid neuron used to predict authenticity.

To reduce the imbalance in the classes, the Synthetic Minority Oversampling Technique (SMOTE) was utilized, and the emotion labels were matched to oversampled embeddings via nearest-neighbor mapping. The model was trained on Adam (learning rate = 1×10^{-4}) with categorical and binary cross-entropy loss and evaluated in terms of accuracy and AUC. This system has compared the generalization ability of frozen EfficientNet feature representation with shallow dense classifiers on dual-task learning. An overview of the embedding-based architecture is shown in Figure 8, emphasizing the frozen feature extraction and shallow classifier setup.

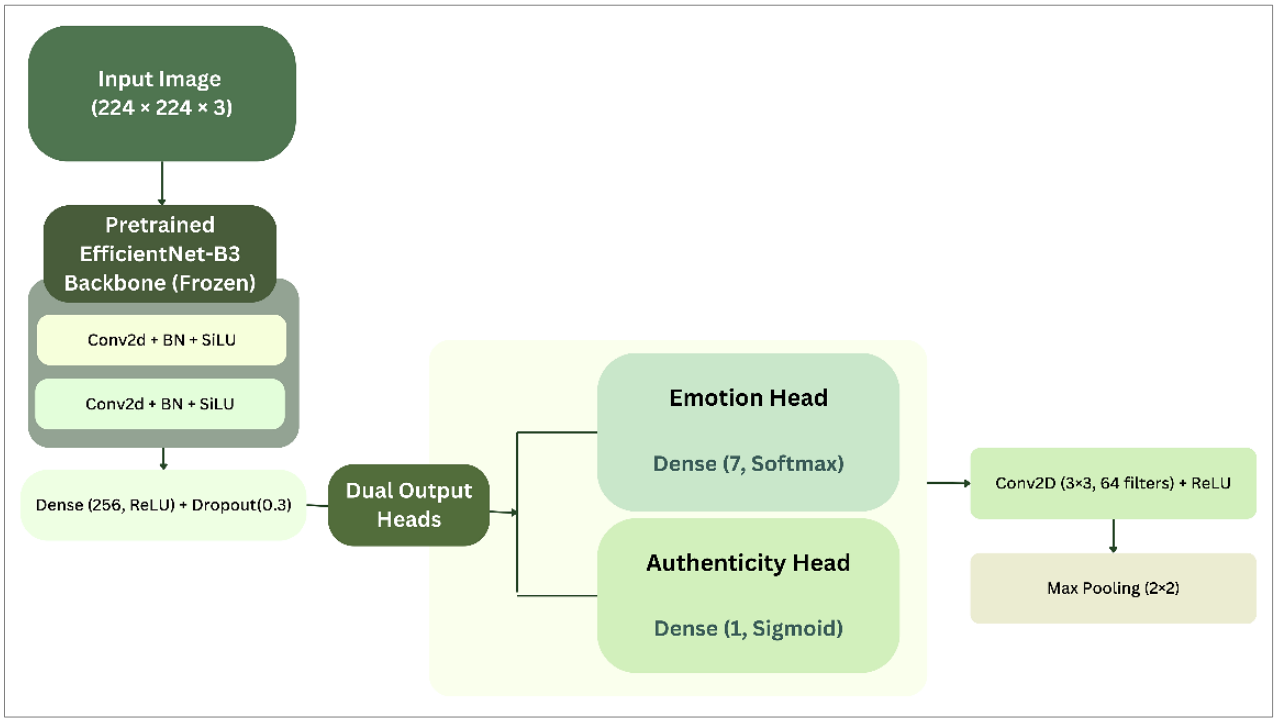


Figure 8. EfficientNet-B3 embedding-based architecture

4-4-4- MobileNetV2 + Xception (Embedding Fusion)

A multimodal embedding fusion framework integrating feature representations from two pretrained convolutional backbones, MobileNetV2 and Xception, both initialized on ImageNet and frozen during training. Each 224 × 224 × 3 input was processed through both encoders to generate 1280-dimensional (MobileNetV2) and 2048-dimensional (Xception) global average pooled embeddings. These were concatenated to form a unified 3328-dimensional representation capturing both lightweight and deep spatial hierarchies. The fused embedding was passed through a dense layer of 256 neurons with ReLU activation and a dropout rate of 0.3 to prevent overfitting.

Two output layers followed: a seven-class SoftMax head for emotion recognition and a single-unit sigmoid head for authenticity detection. SMOTE-based resampling was applied to balance authenticity labels, and emotion labels were aligned via index matching. The model was trained using the Adam optimizer (learning rate = 1×10^{-4}) on categorical and binary cross-entropy losses, and the model performance was evaluated using accuracy and AUC statistics. The objective of this setup was to investigate the effect of cross-model feature fusion on the inter-class discrimination in the domains of emotion and authenticity. This configuration is visually described in Figure 9, illustrating the parallel feature fusion and dense integration layers.

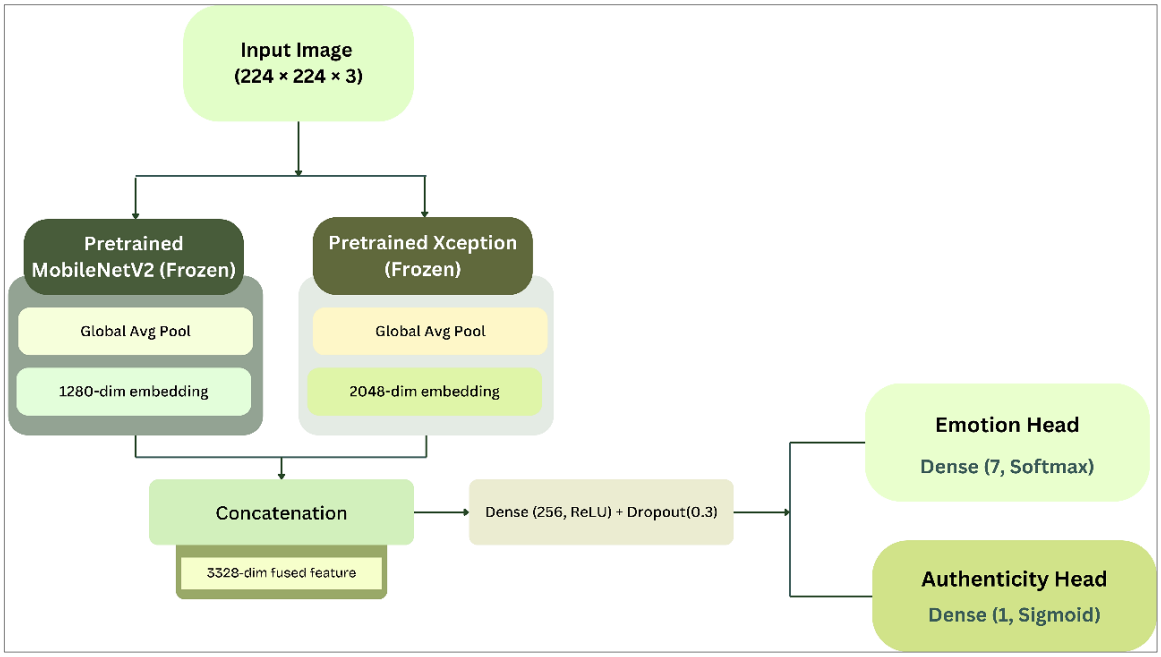


Figure 9. MobileNetV2 + Xception embedding fusion architecture

4-4-5- EfficientNet-B0 (Dual-Branch Without SE)

An efficient two-task framework, which utilizes EfficientNet-B0 as a common backbone to conduct emotion recognition and authenticity prediction at the same time. To obtain shared embeddings, the 224 × 224 × 3 inputs were fed through the pretrained EfficientNet-B0 encoder, which used global average pooling, to obtain shared embeddings. Two task heads were connected two dense-batch-norm-dropout sequence (256 units) with SoftMax activation, which used emotion classification, and a more profound convolution-dense pipeline with sigmoid activation, which used authenticity checking. Training of the network was performed with joint optimization of categorical cross-entropy (emotion) and binary focal loss ($\gamma = 1.5$) with a loss weight of 1.0:1.5. The overall number of parameters was about 11.6 million (11.5 million trainable). This model was used as the ablation baseline to evaluate the effects of SE attention as well as stochastic-depth regularization in the proposed design. The network layout is depicted in Figure 10, showing the dual-branch design before integrating SE attention.

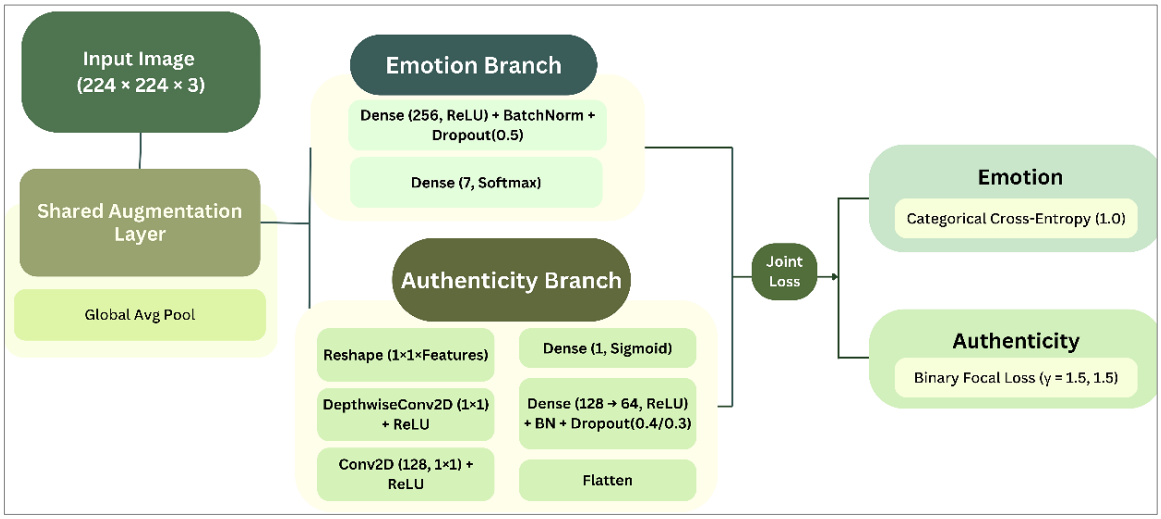


Figure 10. EfficientNet-B0 dual-branch model without SE attention

The suggested dual-task EfficientNet-B0 and SE attention (as described in the previous 4-1) has been used as the baseline configuration to benchmark all the comparative models. All networks were trained and tested with the same set of curated data through stratified sampling to maintain an equal amount of each classification. To ensure fairness, all baseline models were evaluated under comparable training conditions. When pre-trained backbones were used, both frozen and fine-tuned variants were tested, and the best-performing configuration was reported. The proposed model was trained end-to-end under the same optimization constraints, ensuring that performance gains stem from architectural design rather than training advantages.

5- Experimental Results and Discussion

The multi-task framework was meticulously tested on the curated dataset to measure its success in identification of both category-dependent emotions and authenticity cues. The performance was evaluated by accuracy, precision, recall, F1-score, Cohen's Kappa, Matthews Correlation Coefficient (MCC), and AUC. Bootstrapped confidence intervals were presented where appropriate to verify statistical significance.

5-1- Evaluation Metrics

In order to determine the quantitative effectiveness of the emotion recognition and authenticity detection problems, a number of standard evaluation measures were used. These are accuracy, precision, recall, F1-score, Cohen Kappa, Matthews Correlation Coefficient (MCC), and the Area Under the ROC Curve (AUC). The estimation of bootstrapped confidence intervals was also done to provide statistical robustness. These measures are defined as follows:

5-1-1- Accuracy

Accuracy is a percentage of the total correct predictions of all the instances considered. It represents the general performance of the model both in emotion and authenticity tasks. Mathematically it can be represented as:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{2}$$

A higher accuracy indicates better overall model correctness; however, it may not fully capture model reliability in imbalanced datasets.

5-1-2- Precision

Precision (also known as Positive Predictive Value) measures the fraction of correctly identified positive instances among all predicted positives. In authenticity detection, this measures the rate of the falsely identified fake and genuine samples.

$$\Pr e\,cision = \frac{TP}{(TP+FP)} \tag{3}$$

High precision means that there will be a lesser false-positive rate, and that the model is capable of not mistaking one category of data as another one.

5-1-3- Recall (Sensitivity)

Recall is a measure of the capacity of the model to accurately classify all the relevant cases of a particular type. For example, in emotion recognition, it shows the proportion of true happy or angry samples that are correctly identified.

$$Recall = \frac{TP}{(TP+FN)} \tag{4}$$

High value of recall implies that the model does not misleadingly detect real positives but does not account for false alarms.

5-1-4- F1-Score

The F1-Score provides a harmonic mean between precision and recall, offering a balanced measure that is particularly meaningful when dealing with class imbalance, as in the genuine vs. fake subtask.

$$F1 - Score = \frac{(2\times \Pr ecision\times Recall)}{(\Pr ecision+Recall)} \tag{5}$$

A high F1-Score means that classification is consistent and stable across the categories.

5-1-5- Cohen's Kappa (κ)

Cohen's Kappa measures inter-class agreement between predicted and actual labels, adjusted for chance agreement. It gives a more dependable and more conservative assessment of performance than accuracy.

$$\kappa = \frac{(p_o - p_e)}{(1-p_e)} \tag{6}$$

where, p_o denotes the observed agreement and p_e the expected agreement by chance. A value of more than 0.80 implies high reliability and a value between 0.60-0.79 implies a high degree of agreement.

5-1-6- Matthews Correlation Coefficient (MCC)

The MCC is a global measure that takes into account all the four elements of the confusion matrix and is especially informative in the case of binary authenticity detection.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (7)$$

The scale of MCC values is -1 to +1 with +1 being perfect prediction, 0 random performance, and -1 complete disagreement.

5-1-7- Area Under the Curve (AUC)

The AUC, computed from the Receiver Operating Characteristic (ROC) curve, measures the model's discriminative power by evaluating its ability to separate positive and negative classes across all possible thresholds.

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (8)$$

Greater AUC (near to 1.0) implies a great separability and certainty in the authenticity predicted probabilities.

5-1-8- Bootstrapped Confidence Intervals

To guarantee the statistical strength, the results of all significant performance measures: accuracy, F1-Score, and AUC, were averaged with 5 bootstrap resampling runs and 95% confidence intervals were provided where possible. This method alleviates the issue of sample bias and confirms the stability of generalization of the suggested framework.

5-2- Emotion Recognition Results

The initial subtask of the proposed dual-task framework is facial emotion recognition. This test is based on the capability of the model to categorize seven discrete emotion types accurately in a controlled testing condition. The evaluation of performance was based on the use of conventional classification metrics, confusion of analysis and convergence attributes over the training epochs. All classes achieved higher than 0.95 in terms of the F1-score, and Fear, Happy, Surprise obtained a perfect F1-score (i.e., 1.00).

In addition, the Cohen's Kappa (0.982) and MCC (0.982) show high inter class reliability, which confirms that the model is fairly robust to deal with class imbalance.

This high result is explained by the discriminative ability of EfficientNet-B0 backbone that successfully captures hierarchical facial features, as well as SE attention that prioritizes emotionally salient areas. The regulated data collection procedure also minimized the noise associated with pose and illumination allowing better differentiation of emotion categories. All these factors contribute to the high precision and recall rates that have been constantly achieved in all classes.

5-2-1- Summary of Classification Performance

The model performed well for the seven primary emotion classes. From Table 5, we can see that the final accuracy performance is 98.5%, the macro F1-score is 0.983 and the macro F1-score is 0.985.

Table 5. Classification results for Emotion Recognition in the seven categories

Emotion	Precision	Recall	F1-Score	Support
Angry	1.00	0.92	0.96	36
Disgust	0.93	1.00	0.96	40
Fear	1.00	1.00	1.00	59
Happy	1.00	1.00	1.00	60
Sad	0.95	1.00	0.98	40
Surprise	1.00	1.00	1.00	27
Neutral	1.00	0.97	0.99	70
Overall			**0.985**	**332**

Overall F1-Score is calculated by taking the weighted average across all classes.

5-2-2- Confusion Matrix Analysis

The confusion matrix, in Figure 11, of the seven categories of emotions shows the errors in the classification distribution. The majority of categories obtained close-to-perfect discrimination; however, most of the misclassifications happened in Angry (92% recall) and Neutral (97% recall), as they have slight changes from normal facial expressions that are confusable across different categories. This pattern indicates known overlaps in facial muscle activation between low-arousal emotions, indicating that remaining errors are largely perceptual rather than structural model failures.

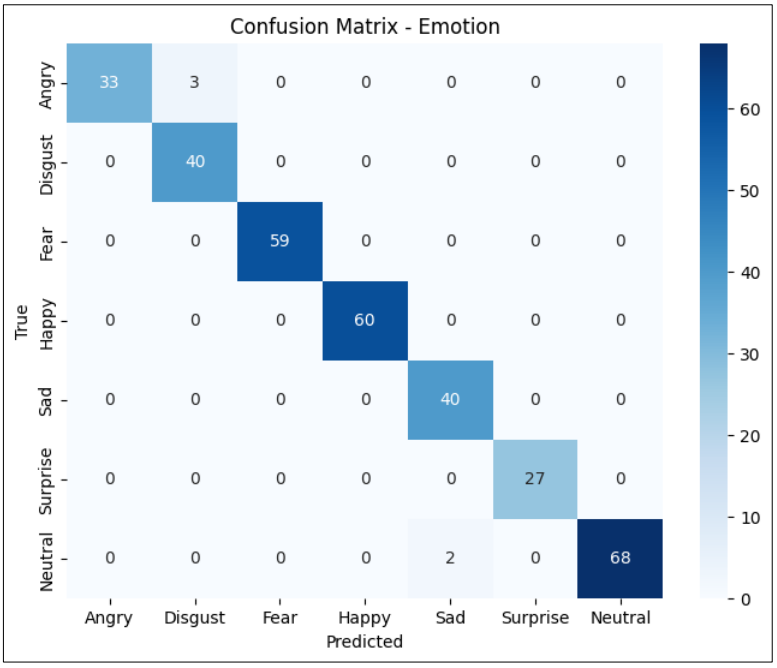


Figure 11. Confusion matrix of seven class emotion recognition task of the proposed model.

5-2-3- Training and Validation Curves Analysis

In order to evaluate the learning stability of the emotion recognition branch, the training and validation accuracy curves are represented in Figure 12. Both curves show a smooth convergence, with the validation accuracy saturating at about 98% and no significant gap between training and validation, indicating minimal overfitting. This consistent trend supports the high ability of the SE-augmented EfficientNet-B0 backbone to generalize, even with moderate data requirements.

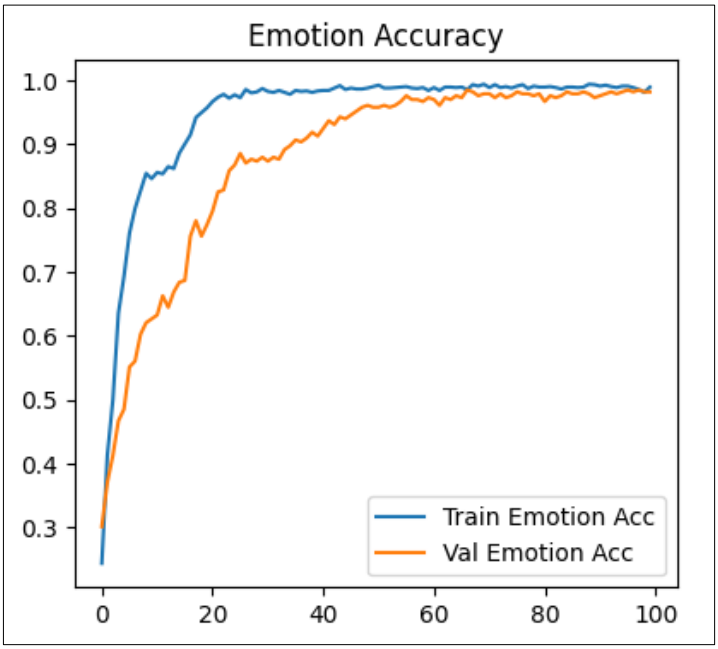


Figure 12. Training and validation accuracy curves for emotion recognition

5-3- Authenticity Detection Results

Authenticity detection, which distinguishes genuine from acted expressions, constitutes the second and more challenging subtask. In contrast to categorical FER, this problem demands a sensitivity to finer, more subtle cues in facial musculature and temporal dynamics. Thus, we assessed the task in terms of not only the traditional accuracy and F1-scores, but also of reliability measures, namely, Cohens Kappa, MCC, and bootstrapped confidence intervals.

5-3-1- Summary of Classification Performance

Authenticity detection was more challenging because of fine-grained differences between authentic and faked expressions. The model obtained an accuracy of 92.2% and a macro F1-score of 0.92 (Table 6). Performance on fake expressions was marginally better (recall = 0.942, F1 = 0.926) than on genuine ones (recall = 0.899, F1 = 0.917), implicate that acted expressions have exaggerated visual cues that the model can more easily capture.

Robustness was also demonstrated as shown by a Cohen's Kappa of 0.843, and MCC factor of 0.843 (with substantial agreement). The bootstrapped assessment achieved an average accuracy of 0.922 with a 95% CI of 0.892–0.949, establishing the statistical validity of the performance being reported.

Authenticity detection is based on finer spatial indicators, as opposed to discrete facial configurations, when compared to emotion recognition. This explains the comparatively lower accuracy, while still demonstrating strong reliability metrics, confirming that the shared emotion-aware representation provides meaningful contextual support for authenticity inference.

Table 6. Classification performance for authenticity detection (genuine vs. fake)

Class	Precision	Recall	F1-Score	Support
Fake	0.91	0.94	0.93	173
Genuine	0.93	0.90	0.92	159
Overall			**0.922**	**332**

Metrics are presented as average values over five bootstrap runs with 95% CI(0.892-0.949).

5-3-2- Confusion Matrix Analysis

The confusion matrix of authenticity detection, as shown in Figure 13, indicates that the model has a high level of discrimination between genuine and fake expressions. Of the total number of fake samples, 163 were rightfully identified, with only 10 falsely identified as genuine. On the same note, 143 actual samples were recognized correctly, and 16 were wrongly classified as fake. It implies that the model is highly accurate and recalls both classes and can capture authenticity cues. The minor instances of misclassification are probably attributed to the nuanced or ambiguous facial expressions that contain similar visual features of the real and fake emotions. These misclassifications are likely caused by borderline expressions where acted emotions closely mimic spontaneous facial dynamics, highlighting the intrinsic difficulty of authenticity assessment from static imagery.

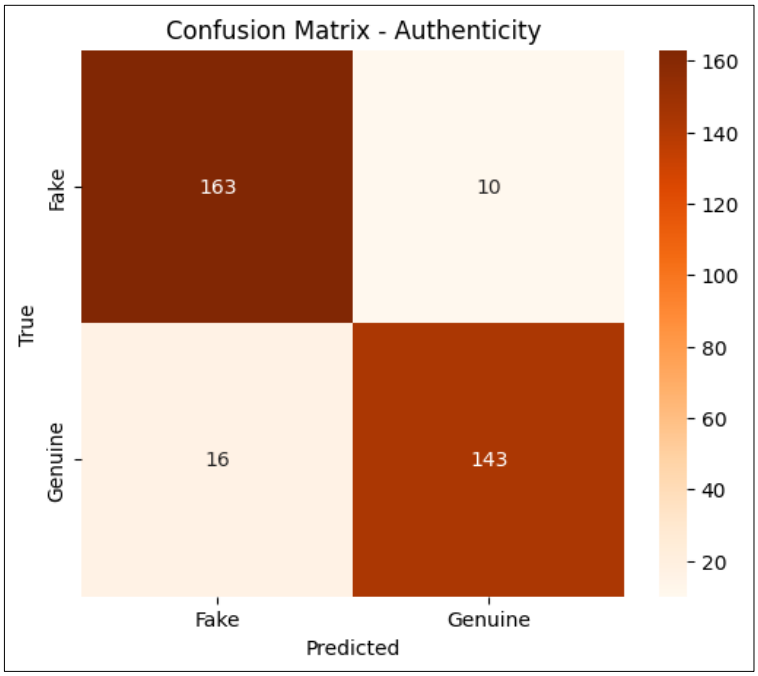


Figure 13. Confusion matrix of authenticity detection (genuine vs. fake)

5-3-3- ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve, as shown in Figure 14, analyzes the relationship between the true-positive and false-positive rates at various thresholds. The authenticity branch scored a 0.922 AUC, indicating that there is high discriminative power between genuine and fake expressions. A high sensitivity and specificity ratio are indicated by the smooth and steeply rising ROC profile toward the upper-left corner, which confirms the credibility of the model confidence estimates.

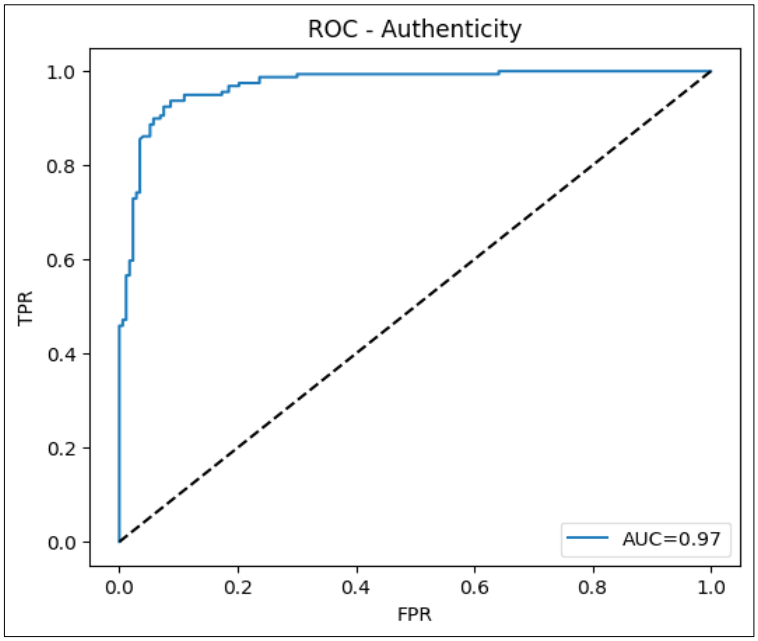


Figure 14. ROC curves for authenticity detection

5-3-4- Precision–Recall Curve Analysis

The discriminative capability of the authenticity detection branch is further confirmed by the Precision-Recall (PR) curve depicted in Figure 15. The model had an Average Precision (AP) of 0.97, which shows that precision is almost ideal in a broad recall threshold. The curve maintains a flat, high-precision plateau before a gradual decline at extreme recall values, signifying balanced sensitivity and reliability even under imbalanced class distributions. These results are consistent with the results of the ROC-based framework and validate the strength of the dual-task framework in authenticity classification and are especially valuable to real-world application, where inaccurate authenticity prediction can be both ethically and practically harmful.

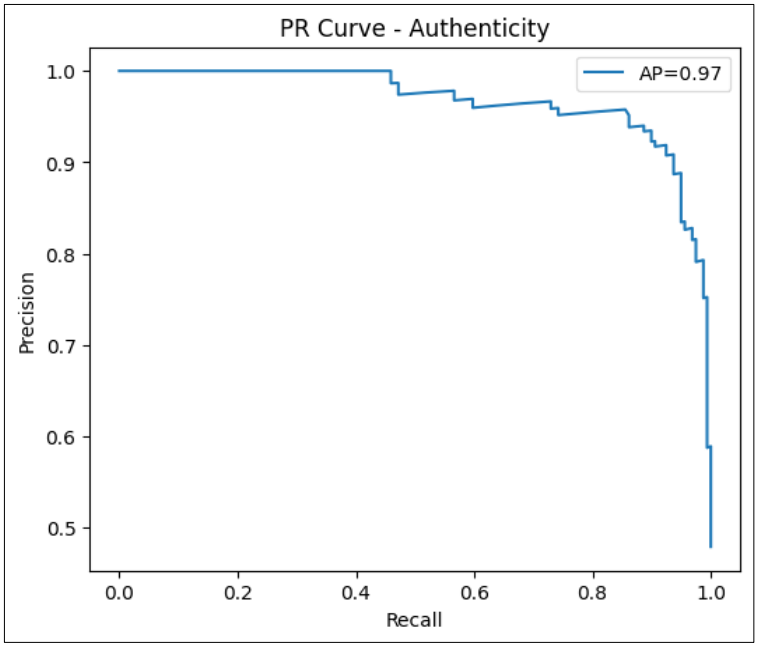


Figure 15. Precision-Recall curve of authenticity detection

5-3-5- Training and Validation Curves Analysis

Figure 16 shows the training and validation curves of the authenticity task, which converge steadily with little variance between the training and validation sets. This behavior indicates the proposed framework has been able to reduce overfitting, although authenticity detection is complex by nature.

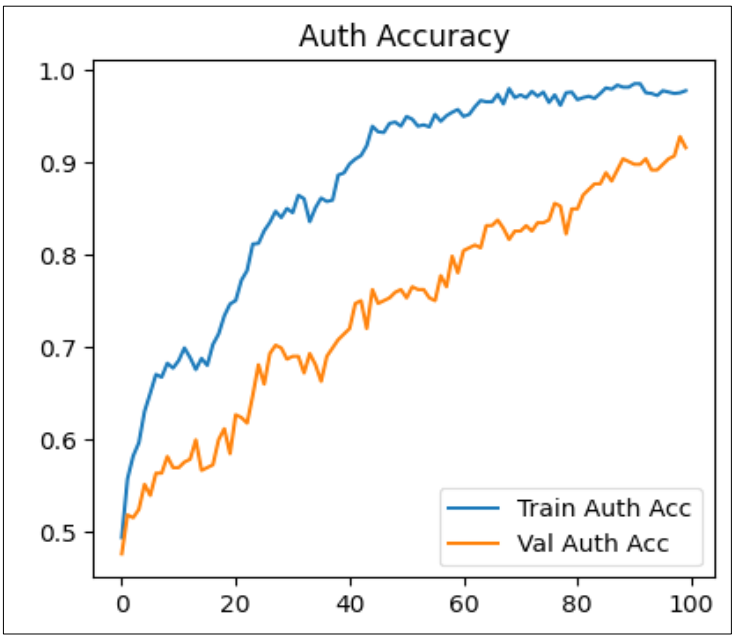


Figure 16. Training and validation accuracy/loss curves for authenticity detection

5-4- Overall Training Dynamics

In order to examine the joint learning behavior of the dual-task model, the combined training and validation accuracy and loss curves are illustrated in Figure 17. The accuracy pattern shows overall improvement in both branches, though with a higher rate of convergence of the emotion task because of the stronger categorical supervision. In the meantime, the authenticity task displays slower yet consistent increases, which can be attributed to the more subtle inter-class differences in the real versus fake expressions.

The associated loss curves show a non-varying monotonically decreasing curve without oscillations, which is indicative of stabilized optimization and lack of divergence. The minimal difference between training and validation losses shows the efficient regularization and strong generalization of the dropout, SE-attention, and stochastic-depth modules. This stability demonstrates the effectiveness of joint loss optimization and SE-based feature recalibration in balancing task dominance during training.

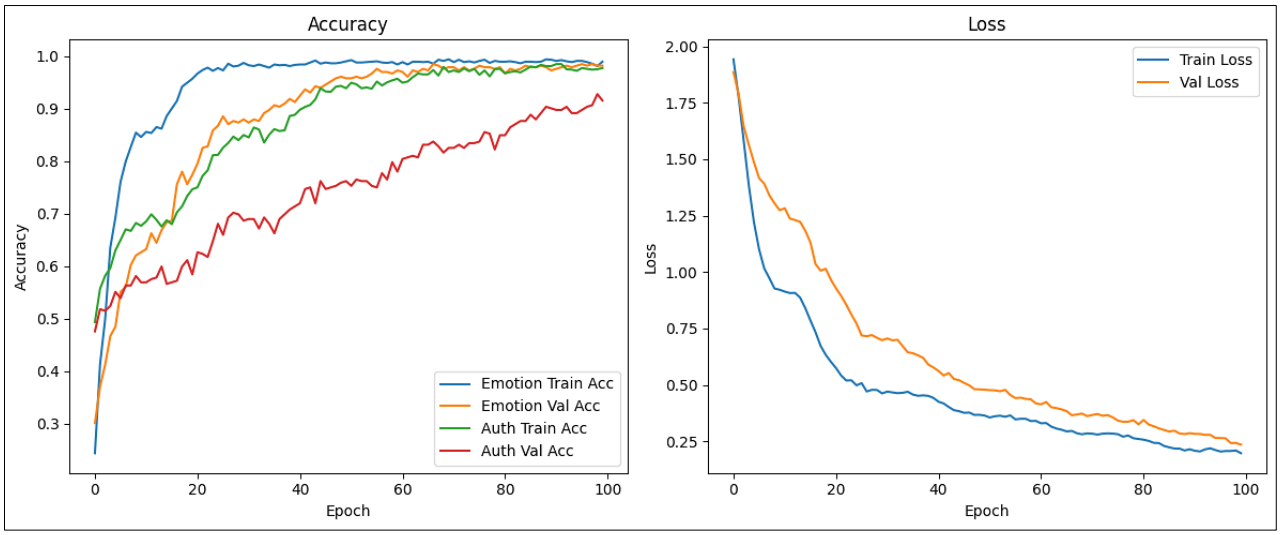


Figure 17. Combined training and validation accuracy and loss curves for both emotion recognition and authenticity detection task

5-5- Comparative Evaluation

To validate the effectiveness of the proposed dual-task framework, we compared it against several baseline and hybrid configurations: CNN, EfficientNetB0 + MobileNetV2, MobileNetV2 + Xception, and EfficientNet-B3. The summarized results in Table 7 and confusion matrices in Figures 18(a) to 18(e) demonstrate the superiority of our proposed model, which achieved 98.5% emotion accuracy and 92.2% authenticity AUC. These results highlight the advantage of attention-enhanced dual-branch optimization to classical architectures.

The dual-task EfficientNet-B0 with attention achieved consistent improvement over all the baselines, especially in the authenticity detection task, where the majority of the single-branch models were challenged by class imbalance and less discriminatory power. However, while MobileNetV2+Xception and EfficientNet-B3 embeddings offered modest gains over a vanilla CNN baseline, the generalization they achieved was limited in comparison with the dual-task approach proposed in this study.

The CNN baseline (Figure 18a) shows many misclassifications between Angry–Sad and Neutral–Disgust. Performance of authenticity prediction is roughly balanced for fake and genuine samples, but separation is limited. The EfficientNet-B0 + MobileNetV2 combinatorial architecture (Figure 18b) provides an emotion recognition accuracy gain while still suffering from relatively high confusion, especially between visually similar emotions, such as Neutral and Sad. In the authenticity task, all samples were biased towards the genuine class, which resulted in high recall and very low discrimination (that is, weak inter-branch feature alignment and poor depiction of authenticity).

The EfficientNet-B3 embedding model (Figure 18c) performs fair with a moderate bias and discrimination among emotions. However, authenticity scores show low levels of sensitivity as frozen backbone embeddings have insufficient representations at the fine-grained level. The MobileNetV2 + Xception fusion (Figure 18d) increases the diversity of feature embeddings that denoising improves the recall rate for Fear and Happy but it cannot be generalized well to Surprise and Sad. There still exists moderate authenticity confusion, some real samples are misclassified as fake (which can be attributed to the redundancy in cross-model fusion).

On the other hand, the dual-branch EfficientNet-B0 architecture without SE (Figure 18e) exhibits significant confusion in the authenticity detection, which verifies that lack of SE attention undermines spatial recalibration effectiveness.

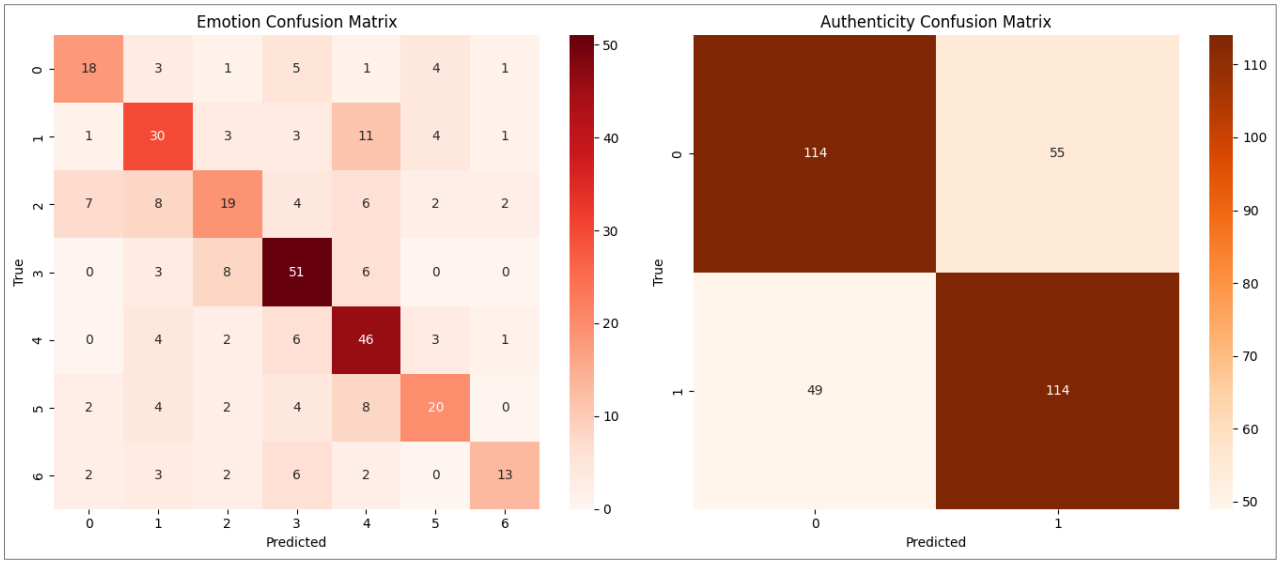
Conversely, the presented dual-branch model was able to minimize these confusions by using squeeze-and-excitation (SE) attention and combined emotion and authenticity branch optimization. This improvement allowed a higher inter-class separability and enhanced generalization when there was an imbalance between the classes.

Overall, these comparisons confirm that neither deeper architectures nor feature fusion alone is sufficient for authenticity modeling. Instead, the combination of dual-task supervision and attention-guided feature refinement is critical for learning discriminative authenticity cues. This explains the consistent superiority of the proposed model across both emotion recognition and authenticity detection tasks.

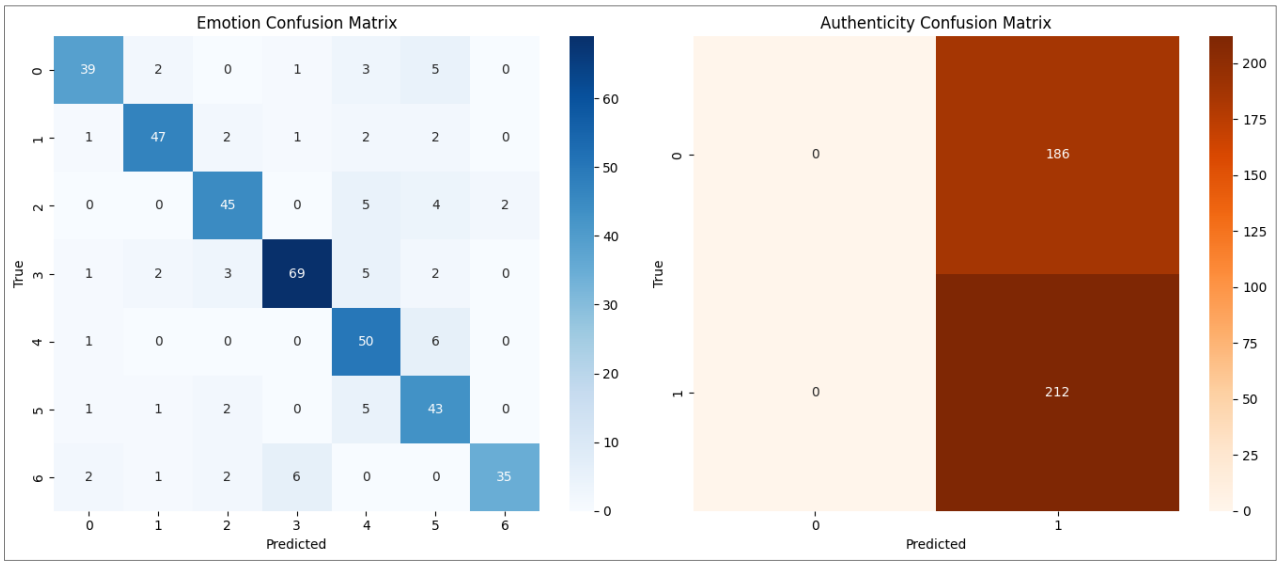
Table 7. Comparison of performance between baseline models and proposed dual-task model

Model	Emotion Accuracy (%)	Emotion F1 (%)	Authenticity AUC (%)	Authenticity F1 (%)
CNN (baseline)	59.0	59.0	69.0	69.0
EfficientNetB0 + MobileNetV2	82.0	82.0	53.0	37.0
EfficientNet-B3 (embed + dense)	62.0	62.0	64.0	63.0
MobileNetV2 + Xception (embedding)	64.0	63.0	67.0	67.0
EfficientNet-B0 (dual-branch)	82.0	82.0	52.0	50.0
Proposed Dual-Task (EfficientNet-B0 + SE Attention)	98.5	98.3	92.2	92.2

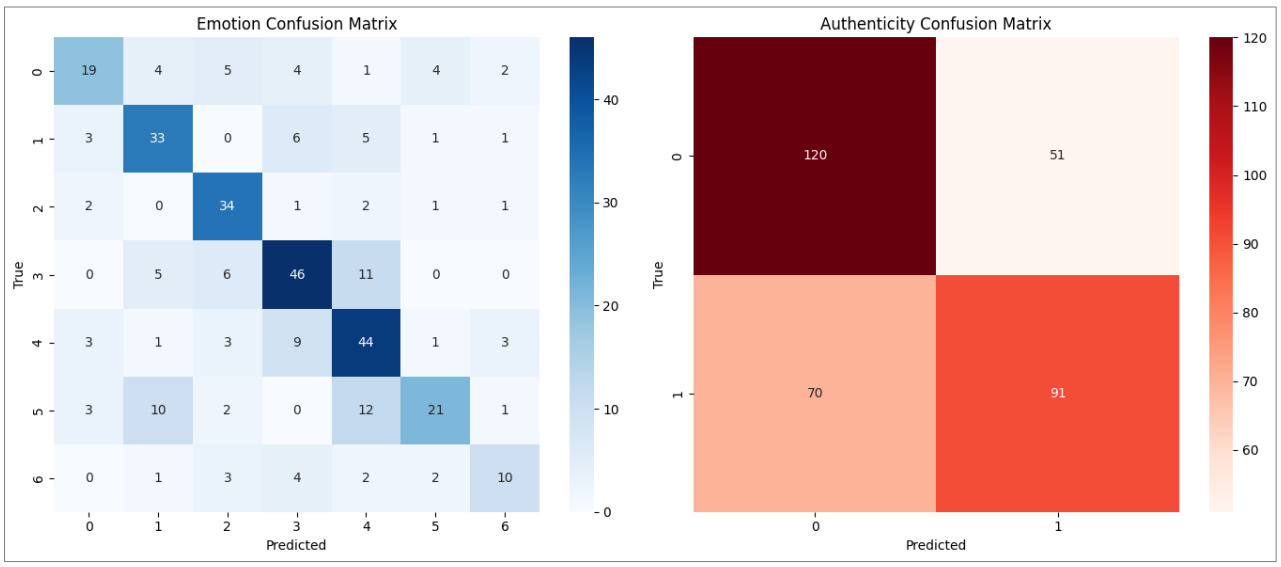
Baseline results are achieved by direct application of classification reports, and the proposed model integrates attention-enhanced branches with joint loss optimization.



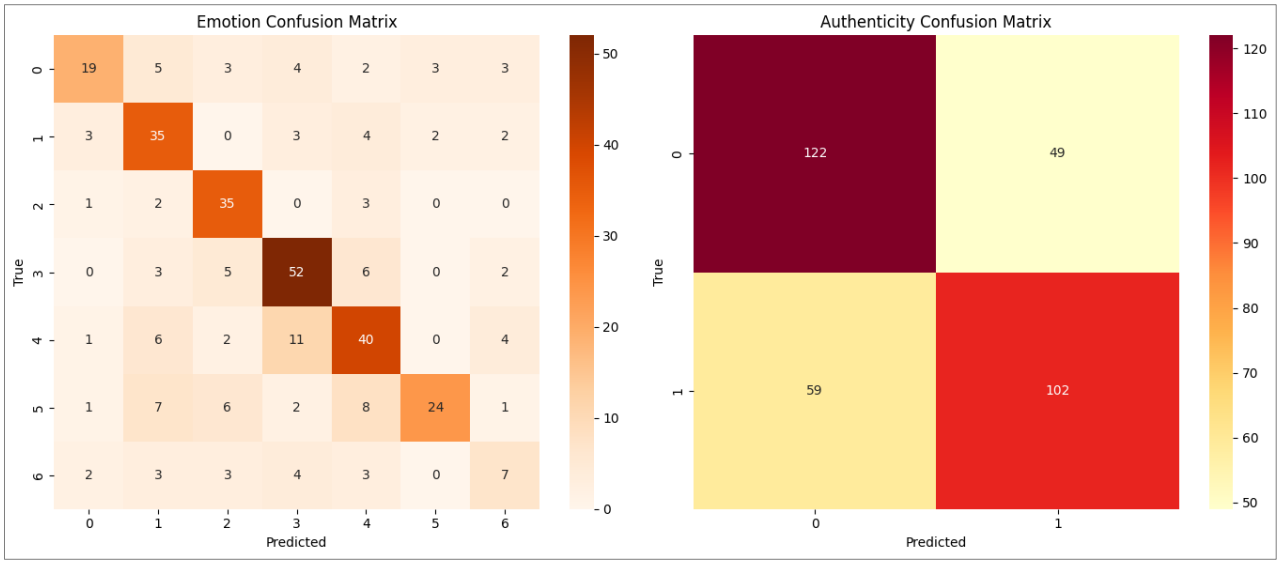
(a) CNN (Baseline)



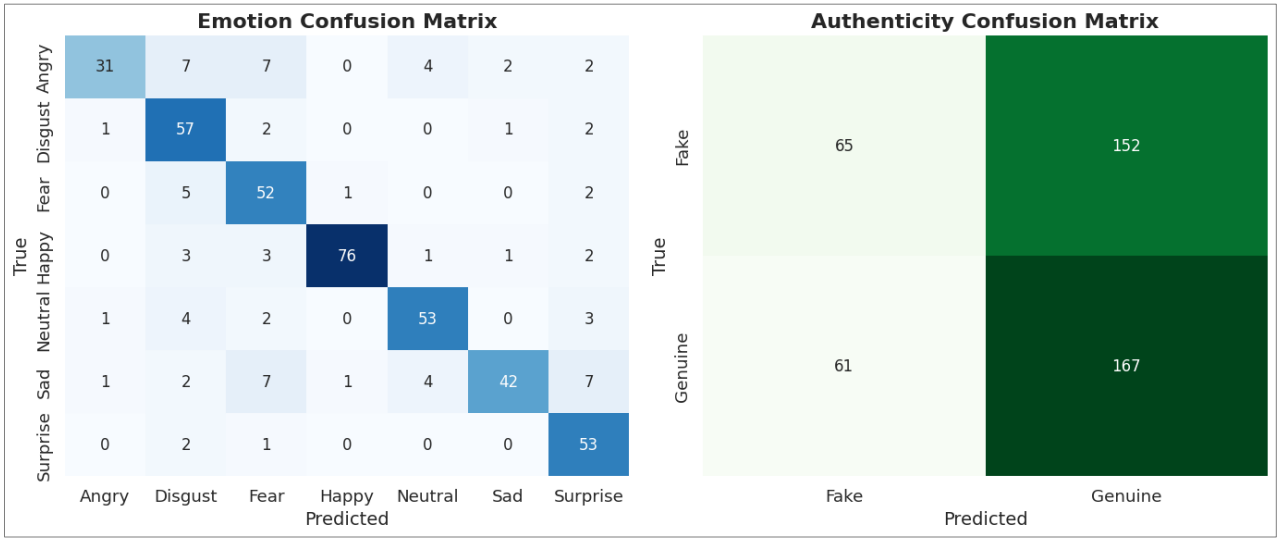
(b) EfficientNet-B0 + MobileNetV2 (Hybrid Branch)



(c) EfficientNet-B3 (Feature Embedding + Dense Classifier)



(d) MobileNetV2 + Xception (Embedding Fusion.)



(e) EfficientNet-B0 (Dual-Branch Without SE)

Figure 18. Comparative confusion matrices of emotion and authenticity recognition across models

5-6- Comparison with Existing Studies

Though there are a number of recent studies that have examined facial emotion recognition and genuine-fake emotion detection separately, there is no direct one-to-one comparison with previous literature because there are no existing frameworks up to our knowledge that can jointly tackle both tasks in a single architecture. Because of this, and in accordance with the previous assessment tradition, the given solution is contrasted individually with the corresponding recent works in the field of emotion recognition and authenticity detection, which are summarized in Table 8.

Table 8. Comparison of the proposed approach with existing facial emotion recognition and authenticity detection studies

Study	Task	Dataset	Model / Approach	Classes	Reported Performance	Limitations
Anand & Babu [11]	Emotion Recognition	FER2013, EMOTIC	EfficientNet-B0 (optimized)	7	96%	Single-task, no authenticity
Manimohan et al. [12]	Emotion Recognition	FER2013, Custom	CNNS, ResNet50	7, 5	79.38%	Sensitive to pose & lighting
Khuntia & Kale [13]	Emotion Recognition	FER2013, FER2013+	CNN, LSTM, ResNet18, and QDA + PCA	7	79.72%	No interpretability
Sunil et al. [23]	Authenticity Detection	ChaLearn, Fake smile master, and Custom	ResNet	Binary	96%	No emotion modeling
Arslan et al. [24]	Authenticity Detection	VREED	Multimodal (ECG + GSR)	4	97.78%	Requires advance hardware
Proposed Method	**Emotion and Authenticity**	**Custom (psychologist validated)**	**EfficientNet-B0 + SE (Dual-Task)**	**7 and Binary**	**98.5% and 92.2%**	**Static images only**

In the domain of facial emotion recognition, contemporary deep learning-based approaches typically report accuracies ranging from approximately 85% to 96%, depending on dataset characteristics, number of emotion categories, and model complexity. Methods employing CNN backbones such as ResNet, MobileNet, and EfficientNet have shown strong performance under both controlled and in-the-wild conditions [11-13]. Compared to these works, the proposed model achieves a higher accuracy of 98.5% and a macro-F1 score of 0.983 across seven emotion classes. This performance gain can be attributed to the incorporation of squeeze-and-excitation attention, which enhances channel-wise feature discrimination, as well as the auxiliary supervision introduced by the authenticity detection task, encouraging the shared representation to focus on psychologically meaningful facial regions.

For authenticity detection, prior studies have largely framed the problem as a standalone binary classification task, often relying on handcrafted features or single-task CNN architectures. Reported accuracies in recent image-based approaches typically fall within the range of 88%–97% [23, 24]. While some video-based or micro-expression–driven methods have achieved incremental improvements, these approaches generally depend on temporal information or specialized datasets that limit practical deployment [18, 19]. In contrast, the proposed framework attains 92.2% accuracy and a macro-F1 score of 0.92 using only static facial images, suggesting that joint learning of emotion categories and authenticity cues enables the model to capture subtle inconsistencies in facial expressions that are less accessible in single-task settings.

It is important to emphasize that variations in datasets, annotation protocols, and evaluation strategies restrict strict numerical comparisons across studies. Most of the existing studies use publicly available datasets having posed expressions or small demographic diversity, but the current study uses a psychologist-validated dataset that was selected to be used in authenticity analysis. In spite of these disparities, the overall improvements that have been reported as compared to the latest standards show that the suggested dual-task model is a significant development. As far as the authors are aware, the present research is one of the first to combine the multi-class emotion recognition and authenticity detection into one attention-enhanced deep learning framework.

5-7- Deployment Validation

The trained dual-task model, which is an extension of EfficientNet-B0 using squeeze-and-excitation (SE) attention, was applied online in a real-time inference pipeline on Render. It was trained using live video input streams and combined face detection, alignment, dual-task inference and result visualization. The pipeline has three major steps:

(i) Detection of a face with the help of the MediaPipe framework provided by Google,

(ii) The dual-task inference to recognize categorical emotions and predict authenticity, and

(iii) Visualization of the results with the help of a web-based interface.

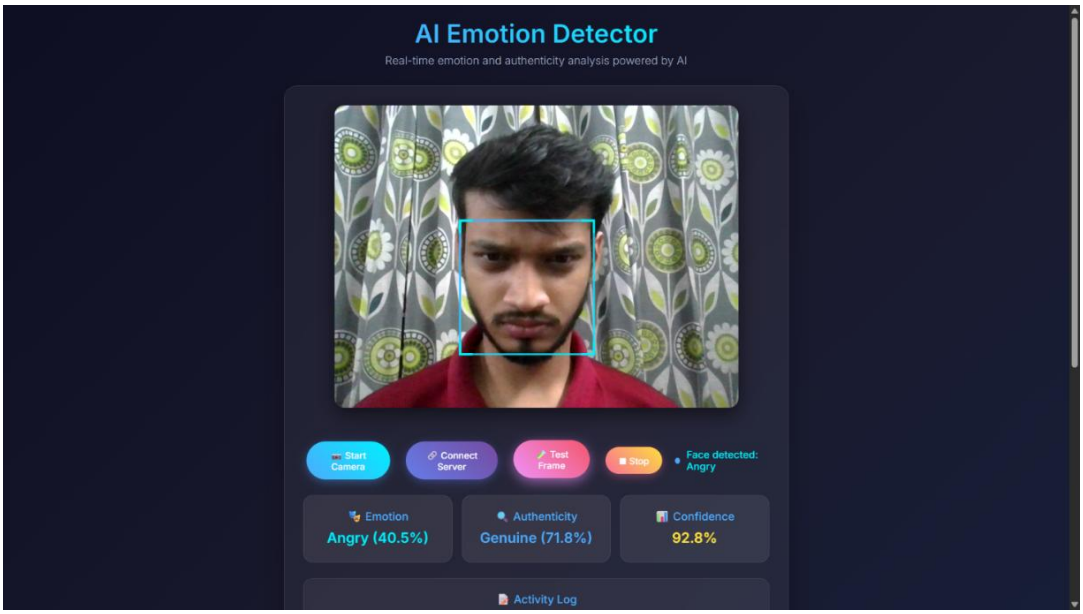


Figure 19. Render hosted real time deployment interface

The deployment in the cloud was based on a CPU-only environment but was able to support an average processing rate of about 5 frames per second, and this was adequate to support smooth interaction over real-time. An image of

the live interface is provided in Figure 19, in which a set of bounding boxes, predicted emotion-based categories, and authenticity labels are overlaid on top of the video feed. This empirical confirmation reveals that the given model can be applied to e-learning analytics, security surveillance, and human-computer interaction.

6- Explainable AI for Model Interpretation

Although deep learning models have a high predictive power, they are frequently considered black-box systems that cannot be easily interpreted. With applications that require emotion and authenticity detection, it is important that there is trust and transparency, especially in sensitive fields like security, healthcare, and human-AI interaction. We used two explainable AI (XAI) methods popular with other scholars (Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME)) to offer post-hoc interpretability of the proposed dual-task model. The two approaches were implemented following model assessment, which ensures that the explanations are based on the real-life decision-making behavior of the trained network.

6-1- Grad-CAM for Visual Explanations

Grad-CAM produces class-discriminative localization maps that indicate the most salient parts of the input that are used to make a particular decision. When used with emotion recognition, Grad-CAM repeatedly prioritized physiologically important locations, including the eyes, mouth, and eyebrows. In the case of authenticity detection, Grad-CAM attention was directed to micro-expressions and peripheral features, which meant that the model was trained to observe small dynamics of the face that differentiate between real and fake expressions.

The resulting heatmaps illustrating these activation patterns are presented in Figure 20, which visualizes the key regions influencing model predictions for both real and fake emotion samples.

Figure 20. Grad-CAM visualizations showing discriminative facial regions for real and fake expressions

6-2- LIME for Local Feature Attribution

Whereas Grad-CAM offers global spatial explanation, LIME supplements this by providing feature importance on local prediction. When applied to detecting authenticity, LIME emphasized fine-textural differences, wrinkles and shading patterns, which affected classification. This view offered by LIME complements Grad-CAM to justify local evidence in order to make decisions of authenticity.

An example of these LIME-based local feature attributions is shown in Figure 21, which highlights pixel-level evidence differentiating genuine and fake facial expressions across emotion categories.

Figure 21. LIME plots of local feature significance between genuine and fake emotion samples

6-3- Complementary Insights and Reliability

Grad-CAM and LIME jointly give explanations that are complementary to each other: Grad-CAM focuses on where the model is paying attention in the face, and LIME shows what local features are causing the model to make predictions. These model explanations are consistent with human-decodable cues (e.g., eyebrows pulling together due to anger, smiling lips on a fake smile) and are consistent with established psychological theories of emotional authenticity. Such a layer of interpretability fosters the trust of the offered system and contributes to its ethical utilization within the real-life context.

7- Conclusion

This paper introduced a dual-task model that integrates the recognition of facial emotions and authenticity detection based on an EfficientNet-B0 backbone with SE attention. The combination of categorical emotion modeling and validity cues modeling by the system provided a close solution to an important limitation in the field of affective computing that exists in the standard single-task FER methods. Built upon an EfficientNet-B0 backbone enhanced with squeeze-and-excitation attention, the proposed model was designed to learn both categorical emotional patterns and subtle authenticity-related cues within a shared representation. A carefully curated dataset, annotated under the supervision of a licensed psychologist, enabled systematic learning of genuine and fake expressions across seven universal emotion categories.

The experiments showed good results: emotion recognition was 98.5% accurate with a macro-F1 score of 0.983, with several emotion classes attaining near-perfect performance and very high inter-class reliability (Cohen's Kappa = 0.982). Authenticity detection, which was inherently more challenging due to subtle expressive differences, had the highest accuracy of 92.2% and a macro-F1 score of 0.92, with substantial agreement (Kappa = 0.843). These findings verify the strength of the shared representation and, at the same time, the challenge of separating the true and fake expressions.

Based on comparative analysis, there were steady improvements across baselines, with the most significant improvement on authenticity classification, which highlights the advantage of learning multiple tasks and integrating attention. In addition to accuracy, explainable AI techniques (Grad-CAM and LIME) showed that the model emphasized semantically relevant areas like the mouth and eyes, which is transparent and is consistent with the evidence of psychology.

Despite these promising results, certain limitations remain. There are constraints such as limited diversity of datasets and lack of time or multimodal information, which could limit the authenticity modeling. Though the dataset was validated by a licensed psychologist, the number of participants (46 subjects) is limited, resulting in incomplete capture of cross-cultural, age-related, and spontaneous expression variations. Therefore, the model is expected to generalize best within controlled or semi-controlled environments. Future directions include larger, culturally diverse, and in-the-wild datasets expected to further improve robustness and external validity. Viable use was also validated through deployment. A pipeline based on FastAPI with MediaPipe detection reached approximately 5 FPS on cloud CPU servers, which can be used to support real-time use in e-learning, surveillance, and human-AI interaction. In short, the suggested dual-task model enhances affective computing by integrating emotion and authenticity with both high precision, interpretability, and deployability, which will be the basis of more transparent and trustful human-AI systems.

8- Declarations

8-1- Author Contributions

Conceptualization, S.T.D., M.J.F., and M.M.R.; methodology, S.T.D., M.J.F., and M.M.R.; software, S.T.D. and M.J.F.; validation, M.M.R., S.A., and O.; formal analysis, M.M.R. and S.A.; investigation, O.; resources, S.T.D. and M.J.F.; data curation, S.T.D., M.J.F., and M.M.R.; writing-original draft preparation, S.T.D. and M.J.F.; writing-review and editing, M.M.R., S.A., Y.S., O., and Z.Z.; visualization, M.M.R., O., and Z.Z.; supervision, M.M.R. and O.; project administration, S.T.D. and M.M.R.; funding acquisition, O.,S.A. and Y.S. All authors have read and agreed to the published version of the manuscript.

8-2- Data Availability Statement

The data presented in this study are openly available in [Mendeley Data] at [doi:10.17632/wmfd4p3z32.1], reference number [34].

8-3- Funding

This study was funded by the Institute for Advanced Research Publication Grant of United International University, Ref. No.: IAR-2026-Pub-013.

8-4- Acknowledgments

We would like to express our sincere gratitude to the Department of Computer Science & Engineering and the Faculty of Science & Information Technology (FSIT), Daffodil International University (DIU), for providing continuous research support and a conducive research environment. We are also thankful to the Department of CSE, United International University (UIU), for their valuable academic support. Additionally, we gratefully acknowledge the psychologist team of Daffodil International University for their assistance in data verification during this research.

8-5- Institutional Review Board Statement

This research was ethically approved by the Department of Computer Science & Engineering, Daffodil International University (DIU). The study was also formally granted approval as a final-year defense thesis by the Faculty of Science & Information Technology (FSIT), DIU. The research project received official ethical and technical approval and was conducted under the institutional reference ID PMS-CSE/DIU-SP25D65423.

8-6- Informed Consent Statement

This study involves the use of facial expression data, which requires capturing human facial images. All participants officially agreed to the use of their facial images as publicly available research data. Written informed consent was obtained from all research participants, and no objections were raised regarding data usage.

8-7- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

9- References

[1] Ullah, S., Ou, J., Xie, Y., & Tian, W. (2024). Facial expression recognition (FER) survey: a vision, architectural elements, and future directions. PeerJ Computer Science, 10, e2024. doi:10.7717/PEERJ-CS.2024.

[2] Agung, E. S., Rifai, A. P., & Wijayanto, T. (2024). Image-based facial emotion recognition using convolutional neural network on emognition dataset. Scientific Reports, 14(1), 14429. doi:10.1038/s41598-024-65276-x.

[3] Kaur, S., & Kulkarni, N. (2024). FERFM: An Enhanced Facial Emotion Recognition System Using Fine-tuned MobileNetV2 Architecture. IETE Journal of Research, 70(4), 3723–3737. doi:10.1080/03772063.2023.2202158.

[4] Scarpazza, C., Gramegna, C., Costa, C., Pezzetta, R., Saetti, M. C., Preti, A. N., Difonzo, T., Zago, S., & Bolognini, N. (2025). The Emotion Authenticity Recognition (EAR) test: normative data of an innovative test using dynamic emotional stimuli to evaluate the ability to recognize the authenticity of emotions expressed by faces. Neurological Sciences, 46(1), 133–145. doi:10.1007/s10072-024-07689-0.

[5] Hasan, M. J., Mohammed, N., Rahman, S., & Koehn, P. (2025). HadaSmileNet: Hadamard fusion of handcrafted and deep-learning features for enhancing facial emotion recognition of genuine smiles. arXiv Preprint, arXiv:2509.18550. doi:10.48550/arXiv.2509.18550.

[6] Sujigarasharma, K., Rathi, R., Visvanathan, P., & Kanchana, R. (2022). Emotion-based human-computer interaction. Multidisciplinary Applications of Deep Learning-Based Artificial Emotional Intelligence, 136–150. doi:10.4018/978-1-6684-5673-6.ch009.

[7] Jia, S., Wang, S., Hu, C., Webster, P. J., & Li, X. (2021). Detection of Genuine and Posed Facial Expressions of Emotion: Databases and Methods. Frontiers in Psychology, 11. doi:10.3389/fpsyg.2020.580287.

[8] Bhagat, D., Vakil, A., Gupta, R. K., & Kumar, A. (2024). Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN). Procedia Computer Science, 235, 2079–2089. doi:10.1016/j.procs.2024.04.197.

[9] Ballesteros, J. A., Ramírez V, G. M., Moreira, F., Solano, A., & Pelaez, C. A. (2024). Facial emotion recognition through artificial intelligence. Frontiers in Computer Science, 6. doi:10.3389/fcomp.2024.1359471.

[10] Ruchita Mathur, & Vaibhav Gupta. (2024). Emotion detection from facial images: A hybrid approach to feature extraction and classification. World Journal of Advanced Research and Reviews, 24(2), 2227–2234. doi:10.30574/wjarr.2024.24.2.3620.

[11] Anand, M., & Babu, S. (2024). Multi-class Facial Emotion Expression Identification Using DL-Based Feature Extraction with Classification Models. International Journal of Computational Intelligence Systems, 17(1), 25. doi:10.1007/s44196-024-00406-x.

[12] P. Manimohan, C.S. Keerthi, Sunkara Kavya Sudha, Devireddy Mourya Chandra Reddy, C. Mahendra, & Raginutala Nagarjuna. (2024). Human Emotion Detection Using CNN and Transfer Learning. International Research Journal on Advanced Engineering and Management, 2(05), 1365–1371. doi:10.47392/irjaem.2024.0188.

[13] Khuntia, A., & Kale, S. (2024). Real time emotion analysis using deep learning for education, entertainment, and beyond. arXiv Preprint, arXiv:2407.04560. doi:10.48550/arXiv.2407.04560.

[14] Islam, U., Mahum, R., AlSalman, A., Sharaf, M., Hassan, H., & Huang, B. (2023). Facial Emotions Detection using an Efficient Neural Architecture Search Network. Researchsquare (Preprint), 1-21. doi:10.21203/rs.3.rs-2526836/v1.

[15] Haider, I., Yang, H. J., Lee, G. S., & Kim, S. H. (2023). Robust Human Face Emotion Classification Using Triplet-Loss-Based Deep CNN Features and SVM. Sensors, 23(10), 4770. doi:10.3390/s23104770.

[16] Singh, A. (2024). Realtime Facial Emotion Detection. International Journal for Research in Applied Science and Engineering Technology, 12(3), 3226–3229. doi:10.22214/ijraset.2024.59630.

[17] Hakim, G. J. P., Simangunsong, G. A., Rangga Wasita Ningrat, Jonathan Cristiano Rabika, Muhammad Rafi' Rusafni, Endang Purnama Giri, & Gema Parasti Mindara. (2024). Real-Time Facial Emotion Detection Application with Image Processing Based on Convolutional Neural Network (CNN). International Journal of Electrical Engineering, Mathematics and Computer Science, 1(4), 27–36. doi:10.62951/ijeemcs.v1i4.123.

[18] Ashraf, A., Gunawan, T. S., Arifin, F., Kartiwi, M., Sophian, A., & Habaebi, M. H. (2023). Enhanced Emotion Recognition in Videos: A Convolutional Neural Network Strategy for Human Facial Expression Detection and Classification. Indonesian Journal of Electrical Engineering and Informatics, 11(1), 286–299. doi:10.52549/ijeei.v11i1.4449.

[19] Pruthviraja, D., Kumar, U. M., Parameswaran, S., Chowdary, V. G., & Bharadwaj, V. (2024). Deep convolutional neural network architecture for facial emotion recognition. PeerJ Computer Science, 10, 1–20. doi:10.7717/peerj-cs.2339.

[20] Cardaioli, M., Miolla, A., Conti, M., Sartori, G., Monaro, M., Scarpazza, C., & Navarin, N. (2022). Face the Truth: Interpretable Emotion Genuineness Detection. Proceedings of the International Joint Conference on Neural Networks, 1–8. doi:10.1109/IJCNN55064.2022.9892298.

[21] Miolla, A., Cardaioli, M., & Scarpazza, C. (2023). Padova Emotional Dataset of Facial Expressions (PEDFE): A unique dataset of genuine and posed emotional facial expressions. Behavior Research Methods, 55(5), 2559–2574. doi:10.3758/s13428-022-01914-4.

[22] Annadurai, S., Arock, M., & Vadivel, A. (2023). Real and fake emotion detection using enhanced boosted support vector machine algorithm. Multimedia Tools and Applications, 82(1), 1333–1353. doi:10.1007/s11042-022-13210-6.

[23] Sunil, M. P., Hariprasad, S. A., Shrishti, S., & Sriharshini, S. (2023). Discrimination Between Fake and Real Emotion Using Modified CNN Model. In Lecture Notes in Networks and Systems: Volume 615, LNNS, 407–416. doi:10.1007/978-981-19-9304-6_38.

[24] Arslan, E. E., Akşahin, M. F., Yilmaz, M., & Ilgın, H. E. (2024). Towards Emotionally Intelligent Virtual Environments: Classifying Emotions through a Biosignal-Based Approach. Applied Sciences (Switzerland), 14(19), 8769. doi:10.3390/app14198769.

[25] Jia, J., Zhang, H., & Liang, J. (2025). Bridging discrete and continuous: A multimodal strategy for complex emotion detection. Proceedings of the 2025 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 1–6. doi:10.1109/MLSP62443.2025.11204253

[26] Govea, J., Navarro, A. M., Sánchez-Viteri, S., & Villegas-Ch, W. (2024). Implementation of deep reinforcement learning models for emotion detection and personalization of learning in hybrid educational environments. Frontiers in Artificial Intelligence, 7. doi:10.3389/frai.2024.1458230.

[27] Evangeline, D., & Parkavi, A. (2024). Facial Emotion Recognition of Online Learners Using a Hybrid Deep Learning Model. International Journal of Intelligent Engineering and Systems, 17(6), 735–751. doi:10.22266/ijies2024.1231.56.

[28] Rathod, V., Chohan, A., Nema, S., Nawghare, A., Devikar, P., & Agrawal, R. (2022). Improved remote mental health illness assessment and detection using facial emotion detection and speech emotion detection. International Journal of Health Sciences, 9577–9590. doi:10.53730/ijhs.v6ns2.7508.

[29] Chethan, R., & Patel, G. L. V. (2024). Deep learning-based emotion detection system. International Journal of Scientific Research in Engineering and Management, 8, 1–13. doi:10.55041/IJSREM36445.

[30] Barnwal, M. A. L., & Barik, M. R. (2025). Human Mood Detection using Image Processing and Machine Learning and Deep Learning. International Journal of Soft Computing and Engineering, 14(6), 28–31. doi:10.35940/ijsce.i9700.14060125.

[31] Zhang, C. (2024). Image-based Facial Emotion Detection SYSTEM. Computer Life, 12(2), 20–26. doi:10.54097/b02twk08.

[32] Singh, M. A. (2024). Real-Time Emotion Recognition System Using Facial Expressions. International Journal of Scientific Research in Engineering and Management, 8(4), 1–5. doi:10.55041/ijsrem31021.

[33] Ton-That, A. H., & Cao, N. T. (2022). Facial Expression Recognition Using a Novel Modeling of Combined Gray Local Binary Pattern. Advances in Human-Computer Interaction, 1–12. doi:10.1155/2022/6798208.

[34] Diya, S. T., Ferdos, M. J., & Rahman, M. M. (2025). Genuine and Fake Facial Emotion Dataset (GFFD-2025). Mendeley Data, 1. doi:10.17632/wmfd4p3z32.1.