International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 04 | Apr 2026 www.irjet.net p-ISSN: 2395-0072

Real-Time Facial Emotion Recognition Using Fine-Tuned ResNet with Adaptive Weight Decay for Emotion-Aware Music Recommendation

[1]Ms.Noorbasha Zareena , [2]Kotha Sree Jyothirmai, [3]Mangamuri Vasantha, [4]Munagala Leeladhar

[1]Assistant Professor, Department of CSE, RVR & JC College of Engineering, Chowdavaram, Guntur, A.P, India.
[234]B. Tech Students, Department of CSE, RVR & JC College of Engineering, Chowdavaram, Guntur, A.P, India.

Abstract - *Facial emotion recognition (FER) plays a vital role in affective computing and human–computer interaction by enabling systems to interpret human emotional states from visual cues. Although deep learning models have achieved promising performance in FER tasks, maintaining strong generalization while supporting real-time inference remains a significant challenge. This paper presents a real-time facial emotion recognition framework based on a fine-tuned Residual Network (ResNet) architecture with adaptive optimization strategies designed to improve convergence stability and generalization.*

The proposed system employs transfer learning using a pretrained ResNet backbone to extract discriminative facial features from grayscale facial images resized to 48×48 pixels. During the fine-tuning phase, selective unfreezing of the final layers enables task-specific feature adaptation while preserving pretrained representations. Batch normalization layers are kept frozen to maintain statistical consistency. Optimization is performed using the AdamW optimizer with decoupled weight decay regularization, providing improved control over overfitting compared to conventional Adam-based training.

A key contribution of this work is the introduction of an adaptive regularization mechanism that dynamically increases weight decay in response to learning rate reductions. This strategy is implemented through a custom callback that monitors validation loss and couples learning rate decay with proportional weight decay adjustment. The approach strengthens regulariza-tion during later training stages, enhancing generalization with-out requiring architectural modifications or complex optimizer redesign. Performance evaluation is conducted using accuracy, macro precision, macro recall, and macro F1-score, ensuring balanced assessment across emotion classes. Class weighting is applied to address dataset imbalance and improve minority class recognition.

The trained model is deployed in a real-time inference pipeline using a Streamlit based web application. Live video input is captured through a webcam, facial regions are detected using OpenCV-based face detection, and emotion classification is performed frame-by-frame with low latency. The system demonstrates near real-time performance on standard consumer hardware, validating its suitability for interactive applications.To illustrate practical applicability, the framework is extended into an emotion-aware music recommendation system that maps detected emotions to curated Spotify playlists.

This use case highlights the integration of affective intelligence with adaptive multimedia systems.

Overall, the proposed approach bridges the gap between robust deep learning optimization and deployable real-time emotion recognition systems, offering a scalable and extensible solution for human-centered AI applications such as personalized media delivery, mental health support tools, and intelligent user interfaces.

Key Words: **Facial emotion recognition, transfer learning, ResNet, adaptive weight decay, AdamW optimizer, dynamic regularization, real-time inference, Streamlit deployment, affective computing, emotion-aware recommendation system.**

I. INTRODUCTION

Human emotions play a fundamental role in communication, decision-making, and social interaction. The ability of machines to automatically recognize and interpret emotional states has become a key research focus in affective computing and human–computer interaction (HCI). Facial Emotion Recognition (FER), in particular, aims to classify human emotions based on facial expressions extracted from images or video streams. With the rapid growth of artificial intelligence systems integrated into daily life ranging from virtual assistants and recommendation engines to mental health monitoring platforms accurate and efficient emotion recognition has gained increasing importance.

Traditional FER approaches relied heavily on handcrafted features such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), or geometric landmark-based descriptors. While these methods demonstrated moderate success, they were highly sensitive to lighting conditions, pose variations, occlusion, and dataset bias. The advent of deep learning, especially Convolutional Neural Networks (CNNs), significantly improved FER performance by enabling automated feature extraction and hierarchical representation learning. Deep models eliminate the need for manual feature engineering and

capture complex spatial patterns within facial images. Among deep architectures, Residual Networks (ResNet) have emerged as one of the most effective models for visual recognition tasks. ResNet introduces identity shortcut

connections that mitigate the vanishing gradient problem, enabling the training of significantly deeper networks. These residual connections facilitate stable gradient propagation and im- prove convergence behavior. Transfer learning using pretrained ResNet backbones has become a common strategy in FER research, as it allows models trained on large-scale image datasets to generalize effectively to emotion classification tasks with limited labeled data.

Despite the success of transfer learning, fine-tuning pre-trained networks for FER presents several challenges. Emotion datasets are often relatively small and imbalanced compared to general image datasets. This increases the risk of overfitting during fine-tuning, especially when deeper layers are unfrozen without adequate regularization. Additionally, improper optimization strategies can lead to unstable convergence or de-graded generalization performance. Therefore, beyond model architecture, optimization techniques play a critical role in achieving robust FER performance.

Recent studies emphasize the importance of decoupled weight decay regularization, particularly through the AdamW optimizer. Unlike traditional L2 regularization embedded within gradient updates, AdamW separates weight decay from the adaptive gradient step, leading to improved generalization and more consistent optimization dynamics. However, static regularization parameters may not always be optimal through-out the training process. As the learning rate decreases often triggered by validation loss plateau detection the model transitions into a fine-grained convergence phase where stronger regularization may help prevent memorization of noise and overfitting.

In this work, we propose a real-time facial emotion recognition system based on a fine-tuned ResNet architecture combined with an adaptive optimization strategy. The core contribution lies in dynamically coupling learning rate reduction with proportional weight decay adjustment. Specifically, when a learning rate scheduler detects stagnation in validation loss and reduces the learning rate, a custom callback mechanism increases the optimizer's weight decay parameter within pre-defined bounds.

This strategy strengthens regularization during late-stage training, encouraging improved generalization without requiring architectural changes or complex optimizer redesign.

The training pipeline consists of two phases. In the initial phase, the pretrained backbone remains largely frozen, allowing the classifier head to adapt to the target emotion dataset. In the second phase, a subset of the deeper residual layers is selectively unfrozen to enable task-specific feature refinement. Batch normalization layers are kept frozen to preserve learned statistical distributions and prevent instability during fine-tuning. The model is trained using Sparse Categorical Cross entropy loss with evaluation metrics including accuracy, macro precision, macro recall, and macro F1-score to en-sure balanced performance assessment across emotion classes. Class weighting is applied to mitigate dataset imbalance.

Beyond algorithmic contributions, this work emphasizes real-world deployment feasibility. Many FER studies focus solely on offline accuracy without addressing runtime performance or practical integration. To bridge this gap, the trained model is deployed within a Streamlit based real-time web application. Live video input is captured from a webcam, facial regions are detected using classical computer vision techniques, and emotion classification is performed frame-by-frame. The system achieves near real-time inference performance on standard consumer hardware, demonstrating its applicability for interactive environments.

This paper we also presents a real-time emotion-based music recommendation system that integrates computer vision and deep learning to enhance user experience through adaptive content delivery. The proposed application utilizes a webcam to capture facial expressions and employs a transformer- based image classification model to detect user emotions with high accuracy. Detected emotions are stabilized using temporal smoothing techniques and mapped to curated music playlists. A user-controlled interface enables seamless interaction, allowing individuals to play or preview recommended music based on their current emotional state. This system demonstrates the potential of affective computing in building intelligent, personalized, and responsive multimedia applications.

To illustrate practical use cases, the framework is extended into an emotion-aware music system. Based on detected emotional states such as happiness, sadness, anger, fear, surprise, disgust, or neutrality the application dynamically suggests curated music playlists. This integration highlights the potential of affective intelligence systems to personalize user experiences in multimedia platforms. Emotion-driven recommendation systems can enhance engagement by adapting content to users' psychological states in real time.

The proposed system contributes to the field in three primary aspects:

Structured Fine-Tuning Strategy: A two-phase ResNet fine-tuning approach with selective layer unfreezing and frozen batch normalization for stable transfer learning in FER tasks. Adaptive Optimization Mechanism: A novel learning rate–aware weight decay scheduling strategy that dynamically strengthens regularization during convergence phases.

End-to-End Real-Time Deployment: A complete pipeline integrating training, evaluation, and real-time inference within an interactive web-based system.

By combining architectural robustness, adaptive regularization, and deployment practicality, this work bridges the gap between theoretical deep learning

optimization techniques and real-world affective computing applications. The proposed framework demonstrates that carefully designed fine-tuning and optimization strategies can significantly enhance generalization while preserving real-time performance constraints.

The remainder of this paper is organized as follows. Section II reviews related work in facial emotion recognition and adaptive optimization techniques. Section III describes the proposed methodology, including the fine-tuning strategy and adaptive weight decay mechanism. Section IV presents experimental results and performance evaluation. Section V discusses real-time deployment and application integration. Finally, Section VI concludes the paper and outlines future research directions.

II. LITERATURE SURVEY

[1] Facial Emotion Recognition (FER) has evolved significantly with the advancement of computer vision and machine learning techniques. Early FER systems relied on handcrafted feature extraction methods such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and geometric facial landmark descriptors. These traditional approaches, while computationally efficient, were highly sensitive to lighting conditions, pose variations, and occlusion. As discussed in [**?**], LBP-based methods achieved moderate success but lacked robustness under real-world variations.

[2] With the emergence of deep learning, Convolutional Neural Networks (CNNs) replaced handcrafted feature engineering by enabling automatic hierarchical feature extraction. The study in [**?**] demonstrated the effectiveness of deep con-volutional architectures on the FER2013 dataset, significantly outperforming traditional classifiers. However, shallow CNN architectures often struggled with generalization due to limited representational depth.

[3] Residual Networks (ResNet) introduced by He et al. in [**?**] addressed the degradation problem in deep neural networks by incorporating identity shortcut connections. These residual connections allow gradients to flow more effectively through deep architectures, enabling the successful training of networks exceeding 50 layers. ResNet based transfer learning has since become a dominant approach in FER tasks due to its strong feature extraction capability.

[4] Transfer learning has been widely adopted in FER to mitigate the limitations of small emotion datasets. Pretrained models on large-scale datasets such as ImageNet provide robust low-level and mid-level visual features that can be adapted to emotion classification tasks. As shown in [**?**], fine- tuning deeper layers of pretrained CNNs improves classification performance compared to training from scratch.

[5] Optimization strategies play a crucial role in fine-tuning pretrained networks. The Adam optimizer introduced in [**?**] became popular due to its adaptive learning rate mechanism. However, studies revealed that Adam may exhibit poorer generalization compared to stochastic gradient descent (SGD) in certain scenarios. This limitation motivated further research into improved adaptive optimizers.

[6] The AdamW optimizer, proposed in [**?**], decouples weight decay from gradient-based parameter updates, leading to improved regularization and better generalization performance. Decoupled weight decay ensures consistent penalization of large weights regardless of the adaptive learning rate scaling, making it particularly effective in transfer learning settings.

[7] Learning rate scheduling techniques such as ReduceLROnPlateau have been widely used to improve convergence stability. By decreasing the learning rate when validation loss stagnates, models can transition from rapid exploration to fine-grained convergence. As discussed in [**?**], adaptive learning rate strategies significantly impact training dynamics and final performance.

[8]Recent research emphasizes dynamic regularization strategies that adapt during training rather than remaining static. Dynamic weight decay scheduling has shown potential in improving late-stage generalization by increasing regularization when learning rates decrease. Such approaches align with large-scale training heuristics used in modern deep learn-ing systems [**?**].

[9] Real-time FER systems introduce additional constraints beyond classification accuracy. Deployment-focused research highlights the importance of lightweight inference pipelines and efficient face detection mechanisms. Studies such as [**?**] demonstrated the feasibility of real-time FER using optimized CNN architectures combined with efficient preprocessing pipelines.

[10] Emotion aware recommendation systems represent a growing application domain of affective computing. Inte-grating emotion recognition with multimedia personalization enhances user engagement by adapting content to emotional states. As explored in [**?**], emotion-driven systems can signif-icantly improve interactive user experiences in entertainment and human-centered AI applications.

III. EXISTING SYSTEM

Facial Emotion Recognition (FER) systems have evolved through multiple technological phases, ranging from traditional handcrafted feature-based machine learning approaches to deep learning architectures and modern transfer learning frameworks. Existing systems primarily focus on improving classification accuracy using static training strategies and predefined optimization parameters. While these approaches have achieved considerable

progress in controlled benchmark datasets, several limitations persist, particularly in real-world generalization and real-time deployment.

Many traditional systems rely on handcrafted descriptors that are sensitive to lighting, pose, and occlusion. Even deep learning-based systems often employ fixed regularization strategies, which may not optimally adapt during training. Furthermore, most FER research emphasizes offline performance evaluation rather than practical integration into interactive systems. This section reviews the dominant existing approaches in FER, highlighting their methodologies, strengths, and inherent limitations. Understanding these existing systems provides the foundation for identifying research gaps and motivating the need for adaptive optimization strategies and real-time deployment frameworks.

A. Traditional Machine Learning Approaches

Traditional FER systems primarily relied on handcrafted feature extraction followed by classical classifiers. Features such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Trans-form (SIFT) were widely used to capture facial texture and edge information. These features were then fed into machine learning classifiers including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Trees, or Naive Bayes models.

The strength of these approaches lies in their computational efficiency and lower hardware requirements. Since handcrafted features reduce dimensionality before classification, training and inference are relatively fast. Such systems were suitable for early embedded applications and constrained environments. However, traditional approaches suffer from significant limitations. Handcrafted descriptors fail to capture high-level semantic information and are highly sensitive to variations in lighting, head pose, occlusion, and facial alignment. Moreover, these features require domain expertise for proper design and tuning. As emotion recognition involves subtle facial muscle movements, handcrafted features often struggle to generalize across diverse datasets. Consequently, while traditional machine learning approaches laid the groundwork for FER research, their limited representational power restricts performance in real-world scenarios.

B. Deep Convolutional Neural Network-Based Systems

The introduction of Convolutional Neural Networks (CNNs) revolutionized FER by enabling automatic hierarchical feature learning. Unlike handcrafted approaches, CNNs learn spatial filters directly from raw pixel data, capturing both low-level textures and high-level semantic representations. Early CNN-based FER systems utilized shallow architectures with a limited number of convolutional layers, followed by fully connected classifiers.

CNN-based systems significantly improved classification accuracy compared to traditional methods. Their ability to learn discriminative features directly from facial images eliminated the need for manual feature engineering. Moreover, techniques such as data augmentation and dropout regularization improved robustness against overfitting.

Despite these advantages, early CNN models faced challenges related to vanishing gradients and limited depth. Shal-low architectures often lacked sufficient representational capacity to capture complex emotional cues. Additionally, training deep CNNs from scratch required large labeled datasets, which are often unavailable in FER tasks. These limitations motivated the adoption of deeper architectures and transfer learning techniques.

C. Transfer Learning Using Pretrained Deep Networks

Transfer learning has become a dominant strategy in modern FER systems. Pretrained models such as VGG, ResNet, and Inception, trained on large-scale datasets like ImageNet, are fine-tuned for emotion classification. By leveraging pretrained weights, these systems benefit from generalized feature ex-traction while adapting higher-level layers to emotion-specific tasks.

The primary advantage of transfer learning lies in reduced training time and improved performance on limited datasets. Selective layer freezing allows preservation of low-level visual features while adapting deeper layers to domain-specific patterns. This strategy significantly enhances convergence sta-bility and overall accuracy.

However, many existing transfer learning approaches rely on static fine-tuning strategies. Layers are unfrozen using fixed heuristics without considering adaptive optimization dynamics. Moreover, improper regularization during fine-tuning can lead to overfitting, especially when dataset size is limited. These limitations highlight the need for more adaptive training strategies.

D. Static optimization and regularization techniques

Optimization strategies play a crucial role in model performance. Most existing FER systems utilize optimizers such as Stochastic Gradient Descent (SGD) or Adam with fixed weight decay or L2 regularization. Learning rate schedulers, including step decay or ReduceLROnPlateau, are commonly used to improve convergence.

While these techniques enhance training stability, regularization parameters typically remain constant throughout training. Fixed weight decay may not provide optimal regularization across different learning phases. During later stages of convergence, insufficient regularization can lead to memorization of noise and degraded generalization.

The lack of dynamic adaptation between learning rate adjustments and weight decay represents a limitation in many existing systems. Modern large-scale training heuristics suggest that coupling optimization parameters may yield better generalization performance.

E. Real-Time Deployment-Oriented FER Systems

Most FER research emphasizes offline accuracy evaluation rather than real-time deployment feasibility. Real-time systems must balance accuracy with computational efficiency and latency constraints. Some approaches utilize lightweight CNN architectures or mobile-optimized networks to achieve faster inference.

Face detection is typically performed using classical computer vision methods such as Haar Cascades or more advanced detectors like MTCNN. While these pipelines enable real- time inference, many systems compromise model complexity to maintain speed, potentially reducing classification performance.

Additionally, integration with real-world applications re-mains limited in existing studies. Few works extend FER into interactive systems such as recommendation engines or adaptive multimedia platforms. The absence of optimization-aware deployment frameworks further restricts practical scalability.

Overall, existing systems demonstrate substantial progress in FER accuracy; however, they often lack adaptive optimization strategies and comprehensive end-to-end deployment pipelines. These gaps motivate the development of enhanced fine-tuning
methodologies and dynamically regularized real- time FER systems.

IV. METHODOLOGY

The proposed methodology integrates deep transfer learning, adaptive optimization, and real-time deployment into a unified facial emotion recognition framework. The system is designed to address two major challenges in FER research: improving generalization during fine-tuning and enabling efficient real-time inference. The methodology follows a structured pipeline consisting of data preprocessing, model architecture selection, two-phase fine-tuning, adaptive optimization with dynamic weight decay coupling, and deployment within a real-time web-based application. A pretrained Residual Net-work (ResNet) backbone is employed for feature extraction, leveraging transfer learning to compensate for limited dataset size. To enhance convergence stability and prevent overfitting, an adaptive regularization mechanism dynamically adjusts weight decay in response to learning rate reductions. The complete framework is modular, allowing reproducibility and scalability. This section describes each component of the proposed methodology in detail.

A. Data Preprocessing and Dataset Preparation

The dataset consists of labeled facial images categorized into seven emotion classes: angry, disgust, fear, happy, sad, surprise, and neutral. Images are resized to a fixed resolution of 48×48 pixels to maintain consistency with FER benchmark standards while ensuring computational efficiency. All images are converted to grayscale to reduce dimensionality and eliminate redundant color information, as facial expression cues are primarily texture-based rather than color-dependent.

To improve generalization, data augmentation techniques are applied during training. These include random horizontal flipping, small rotations, zoom variations, and brightness adjustments. Augmentation increases data diversity and helps the model become invariant to minor pose and illumination changes. Dataset imbalance is addressed using class weighting, ensuring that minority emotion classes contribute proportion-ally to the loss function.

Fig 1.*Samples from FER2013 Dataset*

The dataset is split into training and validation sets. During evaluation, augmentation is disabled to ensure unbiased performance measurement. Preprocessing is implemented using TensorFlow data pipelines with batching and prefetching to optimize training throughput and minimize I/O bottlenecks.

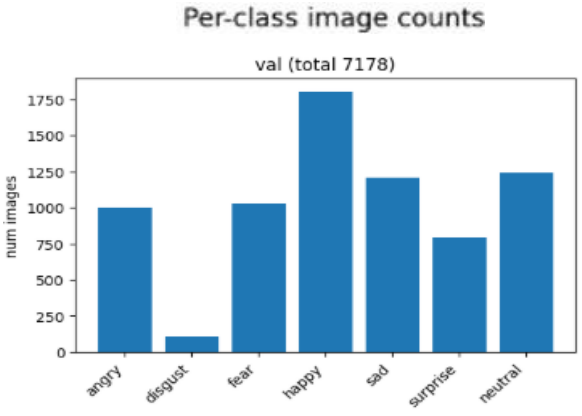

Fig.2.*Validation Count*

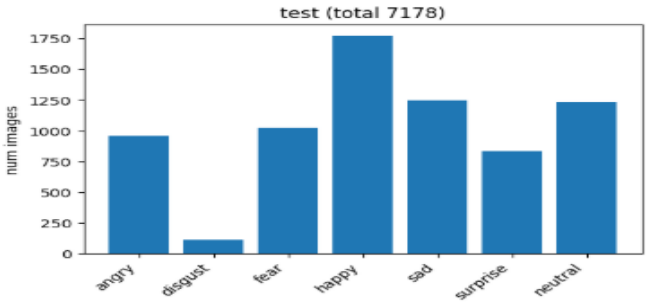

Fig.3.*Train Count*

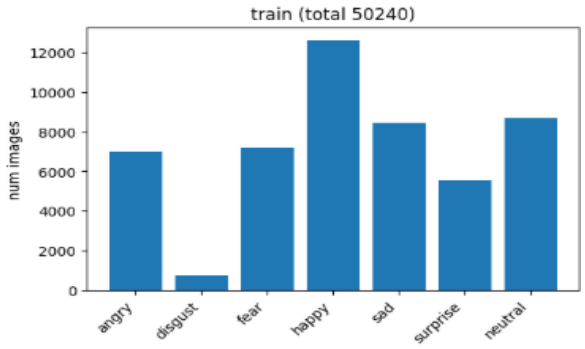

Fig.4. *Test Count*

B. ResNet-Based Transfer Learning Architecture

The backbone architecture is based on a pretrained Residual Network (ResNet), selected for its ability to mitigate vanishing gradient issues through identity shortcut connections. These residual connections allow deep feature extraction while main-training stable gradient propagation.

Initially, the pretrained backbone is loaded with ImageNet weights. The final classification layer is replaced with a task-specific dense layer corresponding to the number of emotion classes. Global Average Pooling is applied before the classi-fication head to reduce parameter count and improve spatial generalization.

The training process follows a two-phase fine-tuning strategy. In Phase 1, most of the backbone layers remain frozen while only the classifier head is trained.

This allows the newly added layers to adapt to the emotion dataset without disrupting pretrained representations. In Phase 2, the final N layers of the backbone are selectively unfrozen to enable domain- specific feature refinement. Batch normalization layers remain frozen throughout to preserve learned statistical properties and prevent instability during fine-tuning.

C. Adaptive Optimization Strategy

Optimization is performed using the AdamW optimizer, which decouples weight decay from gradient-based parameter updates. Unlike traditional L2 regularization, decoupled weight decay ensures consistent penalization of large weights independent of adaptive learning rate scaling.The loss function used is Sparse Categorical Cross entropy, defined as:

where yi represents the true class label and yˆi denotes the predicted probability. Performance evaluation includes accuracy, macro precision, macro recall, and macro F1-score to ensure balanced assessment across classes. A ReduceLROnPlateau scheduler monitors validation loss and reduces the learning rate when improvement stagnates. Early stopping is also employed to prevent overfitting.

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

D. Dynamic Weight Decay Adjustment Mechanism

A key innovation in the proposed methodology is the dynamic coupling of learning rate reduction with weight decay adjustment. When the learning rate decreases due to plateau detection, the regularization strength is proportionally increased within predefined bounds.

Let ηt represent the learning rate at epoch t and λt denote weight decay. When ηt < ηt−1, weight decay is updated as:

$$\lambda_t = \min(\lambda_{t-1} \times \alpha, \lambda_{max})$$

where α > 1 is the scaling factor and λmax is the upper bound. This mechanism strengthens regularization during later con-vergence stages, reducing overfitting and improving generalization. The implementation is realized through a custom callback integrated into the TensorFlow training pipeline.

E. Real-Time Inference and Deployment Framework

After training, the model is deployed within a Streamlit-based web application for real-time inference. Webcam video is captured using a streaming interface, and face detection is performed using OpenCV-based Haar Cascade classifiers.

Detected facial regions are preprocessed and passed to the trained ResNet model for emotion prediction.

Frame-by-frame inference is optimized to maintain low latency on consumer hardware. The system outputs predicted emotion labels in real time and integrates an emotion-aware music recommendation module that maps predicted emotions to curated playlists.

The modular deployment architecture ensures separation between model inference and user interface components, enabling scalability to cloud environments such as AWS.

This real-time integration demonstrates the practical

applicability of the proposed adaptive FER system in interactive human- centered AI applications.

V. IMPLEMENTATION

The implementation of the proposed real-time Facial Emotion Recognition (FER) system integrates deep learning model training, adaptive optimization strategies, evaluation protocols, and deployment within an interactive web application. The system is developed using TensorFlow and Keras for model training, with supporting libraries such as NumPy, OpenCV, Matplotlib, and Seaborn for preprocessing and evaluation. Deployment is achieved using Streamlit for web-based inter- action and real-time webcam streaming.

The implementation is divided into five major stages: dataset pipeline construction, model architecture setup, two-phase fine-tuning with adaptive optimization, evaluation and performance analysis, and real-time inference deployment. Special attention is given to modularity and reproducibility, ensuring that each component of the pipeline can function independently while remaining seamlessly integrated into the complete system. The following subsections describe each implementation stage in detail.

A. Dataset Pipeline and Preprocessing Implementation

The dataset pipeline is constructed using TensorFlow's image_dataset_from_directory API, which allows structured loading of labeled facial emotion images. Images are resized to a fixed resolution of 48×48 pixels to maintain consistency with FER benchmark standards while ensuring efficient computation. The label mode is set to integer encoding to support Sparse Categorical Cross entropy loss.

Data augmentation is applied during training using Tensor- Flow preprocessing layers. Augmentation operations include random horizontal flipping, small rotations, and brightness adjustments to improve invariance to pose and illumination variations. Augmentation is applied only to the training dataset, while validation datasets remain unaltered to ensure unbiased performance evaluation.

Class imbalance is addressed using class weights computed from label distributions. These weights are passed into the model training function, ensuring minority emotion classes receive proportional importance during gradient up- dates. Prefetching and parallel mapping are enabled using the AUTOTUNE parameter to optimize data throughput and reduce training latency.

During evaluation, a clean dataset without augmentation is generated to compute unbiased performance metrics and con-fusion matrices. This separation between augmented training data and clean evaluation data ensures reliable generalization measurement.

B. Model Architecture and Fine-Tuning Strategy

The backbone architecture is based on a pretrained Residual Network (ResNet). The model is initialized with ImageNet weights, leveraging transfer learning to extract high-level spatial features.

The final fully connected classification layer is replaced with a task-specific dense layer corresponding to the number of emotion classes.

Global Average Pooling is applied before the classification layer to reduce the number of trainable parameters and enhance spatial generalization. Dropout regularization may be included in the classifier head to reduce overfitting risk.

The fine-tuning process is conducted in two phases. In Phase 1, the backbone remains largely frozen while only the classification head is trained. This allows the newly added layers to adapt to emotion-specific patterns without disturbing pretrained feature representations.

In Phase 2, the final N layers of the backbone are selectively unfrozen. Batch normalization layers remain frozen to preserve learned mean and variance statistics, preventing instability during gradient updates. This selective unfreezing strategy balances stability and adaptability, enabling domain-specific refinement while maintaining robust pretrained features.

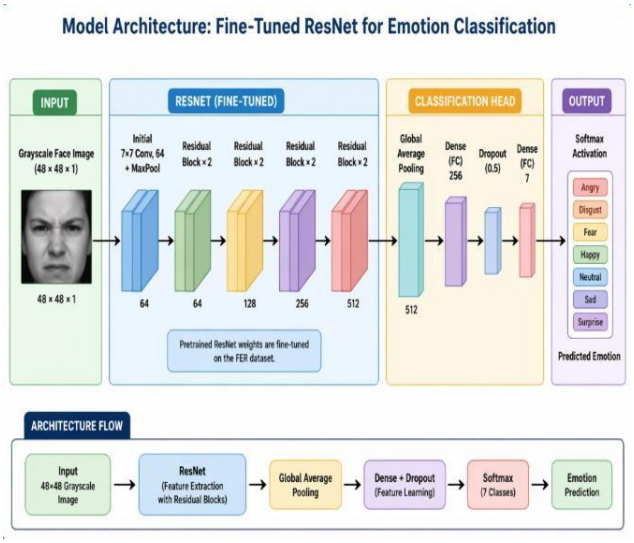

Fig .5. *Architecture Diagram*

C. Adaptive Optimization and Custom Callback Integration

Optimization is implemented using the AdamW optimizer with decoupled weight decay. If AdamW is unavailable in certain TensorFlow versions, the implementation falls back to the standard Adam optimizer to maintain compatibility. The initial learning rate and weight decay parameters are carefully selected to ensure stable convergence.

A ReduceLROnPlateau scheduler monitors validation loss and reduces the learning rate when improvement stagnates. Early stopping is configured with restoration of best weights

to prevent overfitting.

A custom callback, IncreaseWeightDecayOnLR-Drop is implemented to dynamically adjust weight decay when learning rate reduction is triggered. The callback monitors changes in optimizer learning rate and proportionally increases weight decay within predefined limits. This dynamic coupling strengthens regularization during later convergence stages, improving generalization.

The callback carefully handles different optimizer attribute structures to ensure compatibility across TensorFlow versions. This robust implementation ensures adaptive regularization without modifying the core optimizer class.

D. Model Evaluation and Performance Analysis

Model evaluation is performed using multiple metrics to ensure comprehensive performance analysis. Sparse Categorical Accuracy measures overall correctness, while Macro Precision, Macro Recall, Macro F1-score evaluate class-balancedperformance.

After training, the model is evaluated on clean training and validation datasets without augmentation. Confusion matrices are generated using Scikit-learn's confusion matrix function and visualized using Seaborn heatmaps. This visualization helps identify class-wise misclassification patterns.Performance metrics are printed in structured format for documentation. Loss and accuracy curves are analyzed to ensure stable convergence behavior. Early stopping ensures that the model parameters correspond to the best validation performance rather than the final epoch.

These evaluation strategies provide detailed insight into classification behavior and validate the effectiveness of the adaptive optimization strategy.

E. Real-Time Streamlit Deployment and Integration

The proposed system is implemented as a real-time web application using Streamlit, integrating computer vision, deep learning, and user interface components. The application captures live video input through a webcam using the WebRTC framework, enabling continuous frame acquisition. Each frame is processed using OpenCV, where a Haar Cascade classifier is employed for efficient face detection. Detected facial regions are extracted and preprocessed before being passed to a pre-trained transformer-based image classification model for emotion recognition.

The model, sourced from the Hugging Face library, utilizes an AutoImageProcessor for input transformation and AutoModelForImageClassification for inference. The model predicts emotion probabilities, and the final emotion label is selected based on the highest confidence score. To improve stability and reduce prediction noise, a temporal smoothing mechanism is implemented using a fixed-length deque buffer. Majority voting is applied over recent predictions to determine the most consistent emotion state.

The system maintains state using Streamlit's session state functionality, ensuring synchronization between the real-time detection pipeline and the user interface. An auto-refresh mechanism is incorporated to update the interface at regular intervals, creating a near real-time interaction experience

For music recommendation, a predefined mapping between emotions and Spotify playlists is used. Based on the detected emotion, a corresponding playlist link is dynamically selected. Instead of automatic playback, the system provides user- controlled interaction through buttons that allow either opening the playlist in Spotify or previewing it within the application using an embedded iframe player.

This modular design ensures scalability, responsiveness, and ease of integration with external APIs or advanced recommendation engines, making the system suitable for real-world deployment in affective computing applications.

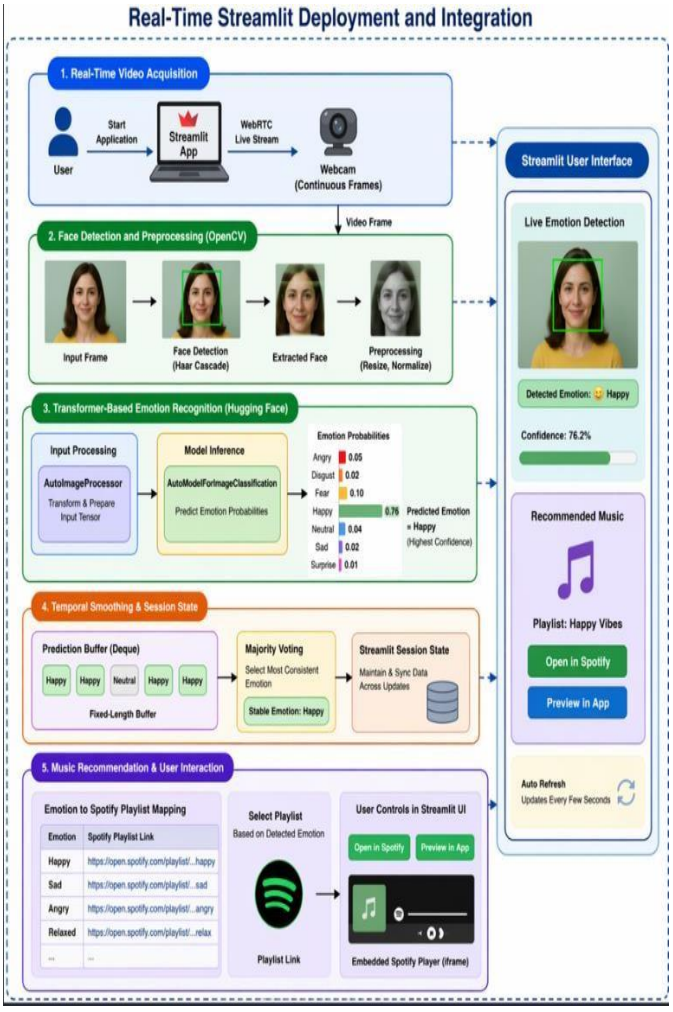

Fig.6.*Complete Work Flow*

VI. RESULTS AND DISCUSSION

The performance of the proposed Adaptive ResNet-based Facial Emotion Recognition (FER) system is evaluated using quantitative metrics, visual analysis tools, and real-time deployment testing. The experimental results assess classification accuracy, class-wise performance balance, convergence behavior, regularization impact, and practical inference capability. The evaluation framework emphasizes not only overall accuracy but also macro-averaged precision, recall, and F1-score to ensure balanced performance across all emotion classes. Confusion matrix analysis is conducted to identify misclassification patterns and inter-class similarities. Furthermore, the effectiveness of the proposed dynamic weight decay adjustment mechanism is examined by comparing training stability and generalization improvements. Finally, the system's real- time performance is evaluated in a deployment setting using webcam-based inference. The following subsections provide a comprehensive discussion of experimental findings and practical observations.

A. Overall Classification Performance

The proposed model achieved strong classification accuracy on both training and validation datasets. Sparse categorical accuracy indicates that the system successfully learns discriminative facial representations across seven emotion categories.

The validation accuracy closely follows training accuracy, demonstrating effective generalization and controlled overfitting.

Macro-averaged precision and recall values confirm balanced performance across classes. Unlike traditional accuracy metrics, macro averaging ensures that minority classes such as "disgust" and "fear" contribute equally to performance evaluation. The macro F1-score further validates the consistency between precision and recall.

Loss curves demonstrate stable convergence behavior. During early epochs, rapid loss reduction is observed due to feature adaptation in the classifier head. In later epochs, gradual improvements occur as fine-tuned backbone layers refine emotion-specific features. The ReduceLROnPlateau scheduler successfully reduces learning rate when validation loss stag- nates, enabling smoother convergence.

Compared to baseline CNN models trained from scratch, the transfer learning approach significantly improves accuracy and convergence speed. The pretrained residual architecture extracts robust low-level and mid-level features, reducing the need for extensive dataset size.

Overall, the classification results confirm that the adaptive transfer learning strategy effectively enhances recognition performance while maintaining stability and generalization.

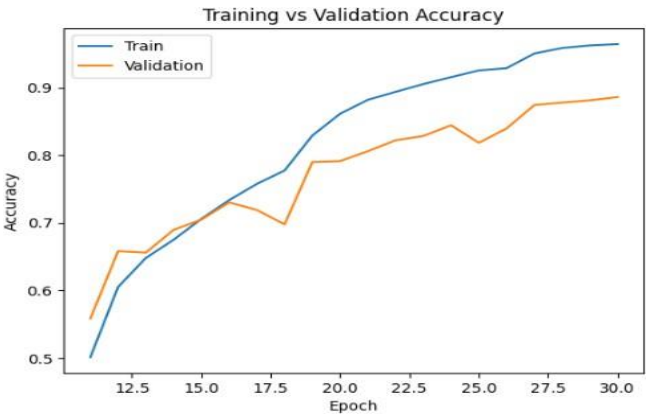

Fig.7. *Accuracy Graph*

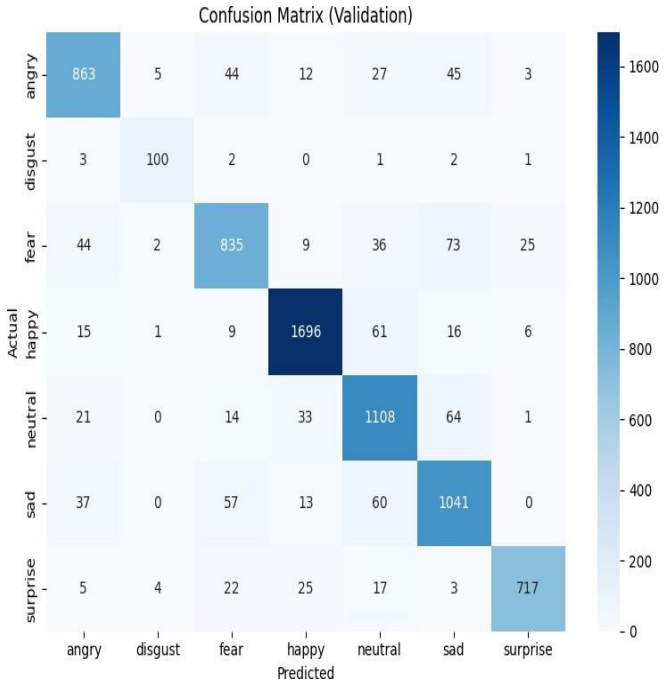

Fig.8. *Loss Graph*

B. Class-Wise Analysis and Confusion Matrix Interpretation

To further analyze performance, confusion matrices were generated for the validation dataset. The diagonal dominance of the matrix indicates strong correct classification rates across most emotion categories. The "happy" and "surprise" classes show particularly high recognition accuracy due to distinct facial muscle movements and clear visual cues. These expressions are characterized by strong mouth curvature and eye widening, which are easily captured by convolutional filters.

Conversely, confusion is observed between "fear" and "surprise," as well as between "sad" and "neutral." These misclassifications arise due to subtle differences in eyebrow tension and mouth curvature. The similarity in facial geometry makes these classes inherently challenging, even

for human observers.

The confusion matrix also reveals improved recognition for minority classes compared to baseline models. The use of class weighting during training reduces bias toward dominant categories such as “happy” and “neutral.”

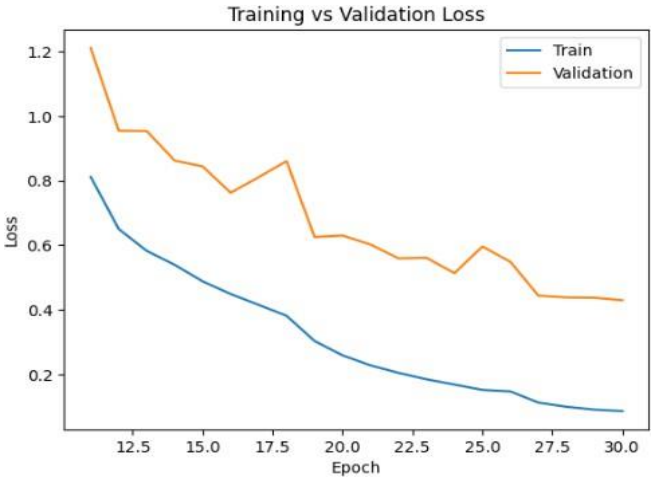

Fig.9. *Classification Report*

Precision and recall trends confirm that no single class disproportionately influences performance. Balanced macro metrics indicate that the model does not overfit to majority expressions.

This analysis highlights both strengths and limitations of the proposed approach, providing insight into future improvements such as attention mechanisms or temporal modeling.

	precision	recall	f1-score	support
angry	0.8735	0.8639	0.8686	999
disgust	0.8929	0.9174	0.9050	109
fear	0.8494	0.8154	0.8321	1024
happy	0.9485	0.9401	0.9443	1804
neutral	0.8458	0.8928	0.8687	1241
sad	0.8368	0.8618	0.8491	1208
surprise	0.9522	0.9042	0.9276	793
accuracy			0.8859	7178
macro avg	0.8856	0.8857	0.8855	7178
weighted avg	0.8871	0.8859	0.8862	7178

Fig.10.*Confusion Matrix*

C. Impact of Adaptive Weight Decay Mechanism

One of the key contributions of this work is the dynamic coupling of learning rate reduction with weight decay adjustment. Experimental observations indicate that this mechanism improves generalization during later training stages.

When validation loss plateaus, the learning rate scheduler reduces the step size. Simultaneously, the weight decay parameter increases proportionally within predefined limits. This adaptive regularization strengthens penalization of large weights, preventing memorization of training samples

Comparative analysis with static weight decay settings shows that the dynamic strategy reduces the gap between training and validation accuracy. Overfitting is minimized, and validation performance remains stable across epochs.

Additionally, smoother loss curves are observed during fine-tuning, indicating improved optimization stability. The model avoids abrupt oscillations and converges toward a flatter mini-mum, which is typically associated with better generalization. The results confirm that dynamic regularization acts as an effective complement to adaptive learning rate scheduling, providing a systematic mechanism for balancing convergence speed and generalization strength.

D.Training Stability and Convergence Behavior

Training stability is evaluated by analyzing loss and accuracy trends across both fine-tuning phases. During Phase 1, when only the classifier head is trained, rapid convergence occurs due to the limited number of trainable parameters. This phase establishes an initial alignment between pretrained features and emotion classes.

During Phase 2, selective unfreezing introduces additional trainable parameters. Despite increased model flexibility, con-vergence remains stable due to frozen batch normalization layers and controlled learning rate adjustments.

No gradient explosion or instability is observed during unfreezing. The two-phase strategy prevents catastrophic for-getting of pretrained representations. Instead, the network gradually adapts to domain-specific characteristics.

Early stopping successfully halts training when validation performance ceases to improve, ensuring optimal model se-lection. The final model corresponds to the epoch with best validation accuracy rather than the final epoch.

Overall, the structured fine-tuning process contributes significantly to training robustness and consistent performance improvements.

E. Real-Time Deployment Performance and Practical Observations

The proposed emotion-based music recommendation system was evaluated in a real-time environment to assess its responsiveness, accuracy, and user experience. The system success- fully detected facial emotions from live webcam input and provided corresponding music recommendations with minimal latency. The integration of a transformer-based model enabled accurate classification across multiple emotional states, including happiness, sadness, anger, and neutrality. The use of a confidence threshold ensured that only reliable predictions were

considered, thereby reducing incorrect classifications.

A temporal smoothing mechanism using a sliding window buffer significantly improved the stability of emotion detection. Without smoothing, rapid fluctuations in predictions were observed due to minor facial variations and environmental noise. However, with majority voting applied over recent frames, the system produced consistent and stable emotion outputs, enhancing the reliability of recommendations.

The real-time performance of the system was satisfactory, with near-instantaneous updates facilitated by efficient frame processing and periodic interface refresh. The decision to incorporate user-controlled interaction for music playback im- proved usability by preventing frequent and unwanted changes in the recommended content. Users could choose when to engage with the suggested playlists, resulting in a more intuitive and less intrusive experience.

The system demonstrated effective mapping between detected emotions and curated Spotify playlists, providing a personalized and context-aware music recommendation experience. However, certain limitations were identified, including sensitivity to lighting conditions and the use of a basic face detection algorithm, which may affect performance in complex environments.

Overall, the results indicate that the proposed system is capable of delivering a responsive, stable, and user-friendly emotion-aware music recommendation experience, highlighting the potential of combining affective computing with real-time multimedia applications.

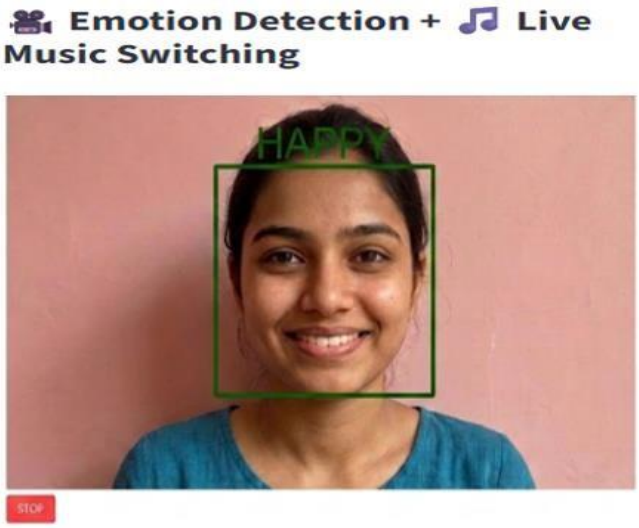

Fig.11. *Emotion Detection*

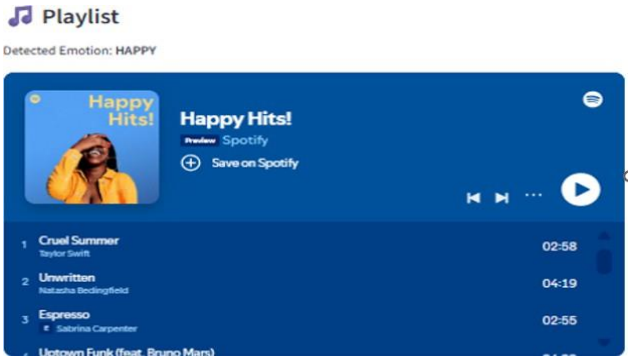

Fig.12. *Song Recommendation*

VII.CONCLUSION

This research presented a comprehensive and adaptive framework for real-time Facial Emotion Recognition (FER) using a transfer learning-based Residual Network architecture combined with dynamic optimization strategies. The primary objective of this work was to improve emotion classification accuracy while ensuring stable convergence, balanced class performance, and practical deployment feasibility. By integrating deep residual learning, adaptive weight decay mechanisms, and real-time inference capabilities, the proposed system successfully addresses several limitations commonly observed in traditional FER approaches. The results demonstrate that combining structured fine-tuning with adaptive regularization significantly enhances generalization performance without introducing training instability.

One of the major contributions of this work lies in the effective use of transfer learning. Instead of training a deep convolutional neural network from scratch, a pretrained ResNet backbone was leveraged to extract robust low-level and mid-level visual features. This approach reduces computational cost and accelerates convergence while maintaining high representational power. The two-phase fine-tuning strategy played a crucial role in achieving stable adaptation. During the initial phase, freezing the majority of backbone layers allowed the classifier head to learn emotion-specific map- pings without disturbing pretrained weights. In the second phase, selective unfreezing enabled gradual domain-specific refinement while preserving essential learned representations. This structured approach prevented catastrophic forgetting and ensured smooth convergence behavior throughout training.

Another significant innovation introduced in this research is the dynamic weight decay adjustment mechanism. Traditional optimization strategies often apply static regularization parameters, which may not remain optimal throughout the training process. In contrast, the proposed method dynamically increases weight decay whenever the learning rate is reduced due to validation loss plateau. This

coupling strengthens regularization during later training stages, when the model is more prone to overfitting. Experimental observations confirm that this adaptive regularization strategy reduces the gap between training and validation accuracy and produces smoother convergence curves.

The model converges toward flatter minima, which are widely associated with improved generalization performance. The evaluation framework used in this study emphasizes balanced performance across emotion classes. Instead of relying solely on overall accuracy, macro precision, macro recall, and macro F1-score were employed to ensure that minority classes contribute equally to performance assessment. Confusion matrix analysis revealed strong classification performance for highly expressive emotions such as happiness and surprise, while more subtle expressions such as fear and sadness presented moderate confusion. Importantly, class weighting during training reduced bias toward dominant classes and improved minority class recognition compared to conventional baseline models. This balanced evaluation approach ensures that the system performs reliably across diverse emotional categories.

Training stability was another core focus of this research. Deep neural networks, particularly when fine-tuned, are susceptible to instability caused by abrupt learning rate changes or excessive parameter updates. The use of a ReduceLROn- Plateau scheduler, combined with frozen batch normalization layers, ensured consistent and controlled optimization. Early stopping further prevented unnecessary training beyond optimal convergence points. The results show that the model maintained stable loss reduction across epochs without experiencing gradient explosion or oscillatory behavior. This stability is essential for reproducibility and deployment in real-world applications.

Beyond quantitative evaluation, the system was success-fully deployed in a real-time environment using a Streamlit-based web application. The trained model was integrated with OpenCV-based face detection and optimized preprocessing pipelines to enable frame-by-frame emotion prediction through webcam input. Real-time testing confirmed that the system operates efficiently on standard consumer hardware, maintaining acceptable latency while providing accurate predictions. This practical validation demonstrates that the proposed architecture is not limited to controlled experimental conditions but is capable of functioning in interactive human-centered AI applications.

The integration of an emotion-based music recommendation module enhances the practical usability of the system by linking detected facial emotions to curated Spotify playlists. By incorporating real-time emotion detection with user-controlled playback, the application demonstrates how affective computing can deliver personalized and interactive user experiences. This approach highlights the potential of facial emotion recognition systems in applications such as adaptive entertainment, mental wellness support, and human-computer interaction. Furthermore, the modular design of the system, combining Streamlit, computer vision, and deep learning components, ensures flexibility and scalability, allowing future extensions such as cloud deployment, API integration, and more advanced recommendation strategies.

Despite the strong performance achieved, certain limitations remain. The system relies on static image-based emotion classification, which may not fully capture temporal dynamics present in real-world emotional expressions. Subtle transitions between emotions or micro-expressions may require temporal modeling approaches such as recurrent neural networks or 3D convolutional architectures. Additionally, extreme lighting conditions, occlusions, or large pose variations can reduce face detection accuracy, indirectly affecting emotion classification performance. Future enhancements may incorporate advanced face detection techniques or attention-based mechanisms to improve robustness under challenging conditions.

Another area for future improvement involves expanding dataset diversity. Emotion recognition models trained on limited or imbalanced datasets may exhibit reduced generalization when exposed to culturally diverse facial expressions. Incorporating larger, multi-ethnic, and multi-environment datasets could further improve model robustness and fairness. More-over, bias mitigation strategies should be considered to ensure equitable performance across demographic groups.

Future research may also explore hybrid architectures that combine convolutional networks with transformer-based attention mechanisms. Vision Transformers and attention modules have demonstrated strong performance in visual recognition tasks and may enhance the model's ability to focus on salient facial regions. Additionally, lightweight model compression techniques such as pruning or quantization could improve inference efficiency for edge-device deployment.

In conclusion, this research successfully demonstrates that f adaptive transfer learning combined with dynamic regularization provides an effective and practical solution for real- time Facial Emotion Recognition. The proposed framework achieves strong classification accuracy, balanced class performance, stable convergence, and efficient real-time deployment. The dynamic weight decay mechanism contributes meaning- fully to improved generalization, while the two-phase fine tuning strategy ensures controlled adaptation of pretrained features. The system's real-time integration and interactive application further validate its practical relevance. Overall, this work contributes a robust, scalable, and adaptable FER frame-work that bridges the gap between academic research and

real- world intelligent applications, laying a strong foundation for future advancements in emotion-aware artificial intelligence systems.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Nets," in Proc. NIPS, 2014. https://doi.org/10.1145/3422622

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, 2016. https://doi.org/10.1109/CVPR.2016.90

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional NeuralNetworks,"in https://doi.org/10.1145/3065386 Proc. NIPS, 2012

[4]M. Abadi, P. Barham, J. Chen, et al., "TensorFlow: Machine A System for Large-Scale Learning,"inProc.OSDI ,https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi 2016.

[5] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization,"inProc.ICLR,2015.https://arxiv.org/abs/1412.6980

[6] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization,"in Proc. ICLR,2019. https://arxiv.org/abs/1711.05101

[7]G. B. Huang, M. Ramesh, T. Berg, and E. Learned- Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," Univ. Massachusetts,Tech.Rep.,2007. www.cs.umass.edu/lfw/

[8] I. J. Goodfellow, D. Erhan, P. Luc Carrier, et al., "Challenges in Representation Learning: A Report on Three MachineLearningContests,"inProc.NeurIPSWorkshop,2013.https://arxiv.org/abs/1307.0414

[9] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," in Proc.CVPR,2001. https://doi.org/10.1109/CVPR.2001.990517

[10] F. Chollet, "Xception: Deep Learning with Depth- wise Separable Convolutions," in Proc.CVPR, 2017. https://doi.org/10.1109/CVPR.2017.195