Design and Implementation of a Multi-Tier Deep Learning Framework for Robust Facial Emotion Recognition Using CNNs, Hybrid Boosting, and Vision Transformers

Ketan Sarvakar [1],
Research Scholar , ksarvakar@gmail.com[1],
Gujarat Technological University, Ahmedabad, India [1]
Dr. Kaushikkumar Rana [2],
Gujarat Technological University, Ahmedabad, India [2]

Received: 17 .03. 2024 **Revised: 18. 08. 2024** **Accepted: 19. 09. 2024**

Abstract

Facial Emotion Recognition (FER) faces challenges such as class imbalance, subtle variations, and limited model generalizability. This paper proposes a three-tier benchmark using the FANE dataset with nine emotion classes. We compare a rule-based Sequential model, five CNN architectures (VGG16, VGG19, ResNet50, InceptionV3, MobileNet), hybrid CNN + Boosting (AdaBoost, GB, XGBoost), and a custom Vision Transformer (ViT), all trained with fixed hyperparameters. Experiments on imbalanced and balanced datasets show that CNN + Boosting performs best post-balancing, while ViT benefits significantly from class balance. Results emphasize the value of standardization and architectural robustness in FER.

Keywords: Facial Emotion Recognition (FER); Class Balancing; Convolutional Neural Networks (CNNs: VGG16, VGG19, ResNet50, InceptionV3, MobileNet); Boosting Algorithms (AdaBoost, Gradient Boosting, XGBoost); Vision Transformer (ViT); Hybrid Deep Learning Models; Fixed Hyperparameters; FANE Dataset

1. Introduction

Facial Emotion Recognition (FER) is important in applications such as health care, monitoring, and interactions between people and computers. However, existing methods of generalization are struggling due to class imbalance, micro-expression differences, and variation in currency and lighting. These challenges are clear in real datasets, especially in Fane, where uneven classes prevent distribution model performance. Traditional CNNs (e.g., VGG16, ResNet50) offer deep feature extraction but often overfit on imbalanced data, failing to detect under-represented emotions. Hybrid approaches that combine CNN with algorithms (eg, Adaboost, XGBoost) improve generalization by focusing on hard samples. Meanwhile, Vision Transformers (ViT), even though they are powerful when it comes to catching global functions, usually support small or unbalanced data sets without adequate balance or pretraining. This article proposes a three-level 9-class measure Fen dataset: (i) a rule-based sequential model such as Baseline, (ii) evaluated with and without increasing CNNs by five, and (iii) trained under a customized base with fixed hyperparameters. We analyze the performance in both unbalanced and balanced datasets to highlight the effect of the preparation of the model on the model's strength and fairness.

Key Contributions:

- A unified FER benchmark comparing shallow, CNN, and ViT models with fixed settings.
- Introduction of CNN + Boosting frameworks evaluated on both imbalanced and balanced FANE datasets.
- Empirical validation of the benefit of dataset balancing, particularly for ViT models.

- Establishment of new performance baselines with class-wise analysis using evaluation metrics.

2. Related Work

Facial Emotion Recognition (FER) has gained traction in domains such as healthcare and HCI. CNNs are most commonly used because of their strong learning capabilities. However, several studies lack stability in datasets, hyperparameters, and training setups, which limits the reproducibility and generalization of their results.

In [1] suggested an EfficientNet-XGBoost model that performed poorly on FER2013 but well on small datasets (CK+, JAFFE). [2] used several CNNs to achieve high CK+ accuracy without resorting to testing generalization. In a similar vein, studies by [3], [4], and [5] employed different CNNs but lacked fixed hyperparameters, making comparisons unfair. [6] used histogram equalization to improve VGGs, but overfitting on FER2013 remained a problem. Classification layers were not included in the scope of [7] hyperparameter tuning.

Hybrid models combining CNNs with boosting remain underexplored. [8] presented CNN + XGBoost fusion, but only on small datasets. ViT-based models were analyzed by [9], while [10] proposed an interpretable multi-branch RBF network yet both lacked hybridization and fixed configurations.

Gaps Identified:

- Poor cross-dataset generalization [[1], [2], [5], [9]]
- No standardized training pipeline [[3] , [4] , [5] , [6] , [7] , [8], [10]]
- Bias due to class imbalance [[1], [3], [6], [10]]
- Limited use of hybrid CNN + Boosting [[1], [8]]
- Inconsistent evaluation across datasets [[1], [2], [6], [8], [10]]

This study addresses these issues by proposing a unified benchmark using five CNNs, three boosting algorithms, and a Vision Transformer, all trained with fixed hyperparameters on both imbalanced and balanced FANE datasets.

3. Methodology

This study follows a three-tiered approach to the use of traditional, CNN-based, and transformer-based models. The experiments are performed on a FEN data set, including 16,912 unbalanced images in 9 emotional classes. To address the square imbalance, data text (horizontal flip, ± 10% shift, ° 20 ° rotation) is used, resulting in a balanced data set per square (26,100 in total) with 2900 images. Images are shaped in 75 × 75 pixels, which are converted to RGB, normalized, and split class-wise: 10% test, 80% train, 20% validation.

- Tier-1 establishes a baseline using a sequential-based rule-rating and achieves 55.56% accuracy of unbalanced data.
- Tier-2 evaluates five CNNS: VGG16, VGG19, Resnet50, Inceptionv3, and MobileNet, which are trained with fixed hyperparameters (Adam, LR = 0.001, batch = 128, epochs = 200). Their built-in is classified using a more advanced algorithm (Adaboost, Gradient Boosting, XGBOOST) to produce hybrid CNN-Boosting models.

- Tier-3 trains a customized Vision Transformer (Vit), Introduction to 75 × 75 images (15 × 15 patches, 768-dim embeddings, 6 encoder blocks, 8 heads, 8 heads, LR = 3e-5, epochs = 25). This allows the fair benchmarking of Vit under similar settings to CNN and hybrid approaches.

This feature ensures frequent evaluation by highlighting the effect of balance and architectural choice on performance.

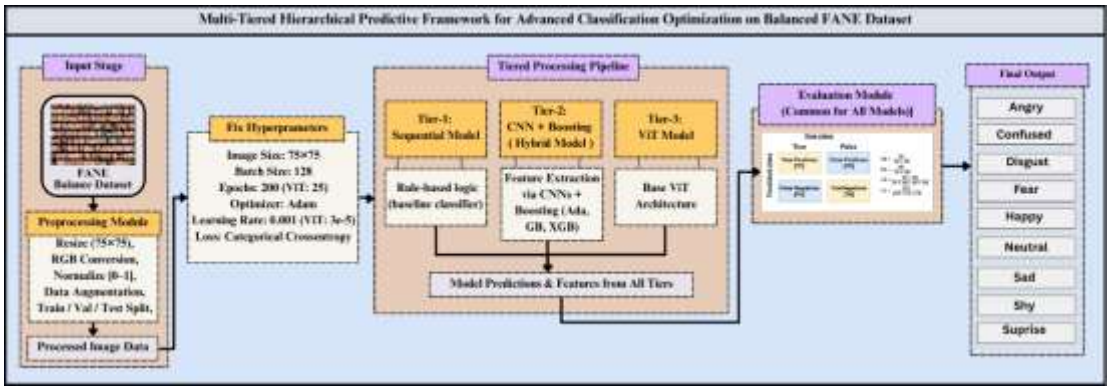


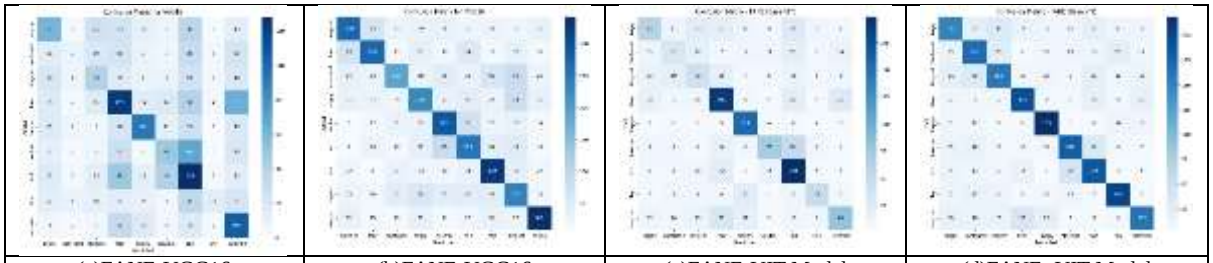
Figure 2. Confusion matrix of Imbalanced and Balanced FANE Dataset Models

```
ALGORITHM Compact Multi-Tiered Predictive Framework (FANE Dataset)
INPUT: FANE_Dataset (Balanced), Hyperparameters
OUTPUT: Predicted_Classes, Metrics

    // Step 1: Preprocess Data
    Processed_Data ← Preprocess(FANE_Dataset, Size=75x75, Normalize=[0-1])
    Train_Data, Test_Data ← TrainTestSplit(Processed_Data, 80:20)

    // Step 2: Tier 1 - Sequential Model
    Sequential_Features ← ExtractSequentialFeatures(Train_Data)
    Sequential_Predictions ← TrainSequentialModel(Sequential_Features,
                                Optimizer="Adam", LR=0.001, Epochs=200, Batch=128)

    // Step 3: Tier 2 - Hybrid CNN + Boosting
    Hybrid_Predictions ← {}
        FOR cnn IN ["VGG16", "MobileNet"] DO
            CNN_Features ← ExtractCNNFeatures(Train_Data, Model=cnn)
            Hybrid_Predictions[cnn] ← TrainHybridModel(CNN_Features, Boost="XGBoost",
                                        Optimizer="Adam", LR=0.001, Epochs=200, Batch=128)
        END FOR

    // Step 4: Tier 3 - Base ViT
    ViT_Predictions ← TrainViTModel(Train_Data, Optimizer="Adam", LR=3e-5, Epochs=25, Batch=128)

    // Step 5: Fuse and Evaluate
    Combined_Predictions ← EnsembleVote([Sequential_Predictions, Hybrid_Predictions, ViT_Predictions])
        FOR each sample IN Test_Data DO
            Predicted_Class ← Predict(Combined_Predictions, sample)
            True_Class ← GetTrueLabel(sample)
            Metrics ← ComputeMetrics(Predicted_Class, True_Class, ["Accuracy"])
        END FOR
    RETURN Predicted_Classes, Metrics

END ALGORITHM
```

4. Experimental Setup & Results

To ensure frequent evaluation, a standardized experimental environment was used, including an Intel Xeon CPU with an Nvidia Tesla P100 GPU (16 GB VRAM) and 64 GB of RAM. Tensorflow 2. X and Keras were hired for workflows with deep learning, while Scikit-Learn and XGBOOST were used for clothing training tasks. All images were resized to 75×75 pixels, converted to RGB, and normalized to [0,1]. The dataset was split into training, validation, and

testing subsets following an 80:10:10 stratified scheme. Each CNN model, VGG16, VGG19, ResNet50, InceptionV3, and MobileNet has been trained using fixed hyperparameters: Adam optimizer (LR = 0.001), batch size = 128, epochs = 200, and categorical cross-entropy as the loss function. To standardize architecture, each model was extended with GAP → Dense(256) → BN → Dropout(0.5) → Softmax layers. From these CNNs, deep features were extracted and then fed into three ensemble classifiers: AdaBoost, Gradient Boosting, and XGBoost.



(a)FANE-VGG16 Imbalance dataset	(b)FANE-VGG16 Balance dataset	(c)FANE-VIT Model Imbalance dataset	(d)FANE- VIT Model Balance dataset

Figure 2. Confusion matrix of Imbalanced and Balanced FANE Dataset Models

Model performance has been evaluated using multiple calculations: accuracy, accurate, recall, F1 score, and RMSE. As shown in Figure 2, confusion for imbalanced datasets indicates severe overfitting. Although the Reset50 + increase, for example, reached 88.89% accuracy, VGG16 was seen as "fear" as "happy" 122 times, to confirm weak generalization.

Tabel 1. Model Performance and Parameters on Balanced FANE Dataset

Model Performance and Parameters on Balanced FANE Dataset			
Models	**Balance Dataset Accuracy**	**Fix Hyperparameters Used**	**Dataset**
BaseVIT	**61.24%**	**Adam optimizer, LR = 3×10^{-5}, Epochs = 25**	**FANE Dataset**
VGG16 + Boosting	**51.49%**	**Adam, LR = 0.001, Epochs = 200, Batch = 128, Img = 75×75**	
VGG19 + Boosting	**47.28%**		
RESNET50 + Boosting	**33.25%**		
InceptionV3 + Boosting	**57.17%**		
MobileNet + Boosting	**63.67%**		

To overcome this, the FANE dataset was rebalanced using augmentation. Following this, all models were retrained under the same fixed hyperparameters. While raw accuracy slightly declined, generalization has notably improved. As per Table 1, MobileNet + Boosting has achieved the highest balanced accuracy (63.67%), outperforming ResNet50 (33.25%) and InceptionV3 (57.17%), which has shown more resilience to class balancing. The Base ViT model is trained separately with low learning rate (3×10^{-5}) and fewer epochs (25), due to its various convergence behaviors. Initially, on unbalanced datasets, it has only achieved 50.38% accuracy, but after being applied to a balanced version, it has improved up to 61.24%. The confusion matrix (Figure 1) shows obvious diagonal dominance, signaling better class prediction stability.

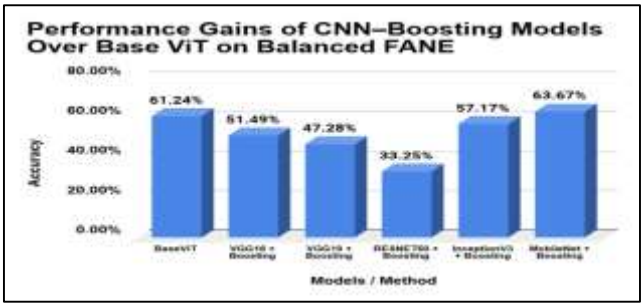


Figure 3. Performance Gains of CNN–Boosting Models Over Base ViT on Balanced FANE

As illustrated in Figure 3, a comparative bar chart has been constructed, confirming that MobileNet + Boosting has surpassed Base ViT by over 2% on the balanced dataset. This result reaffirms that a hybrid CNN–Boosting framework, when paired with dataset balancing and fixed hyperparameter settings, has provided superior FER performance across multiple architectures.

5. Discussion

Experimental conclusions outline the data set balance and the important role of architectural alternatives in Facial emotion Recognition (FER). High accuracy of unbalanced dataset (eg, ResNet50 + 88.89%) was manifested through confusion, which was motivated by overfitting for large sections such as "happy" and "neutral", such as under-paired emotions such as "fear" and "hate". This behavior seriously limits the gratitude of the real world. When you use square growth to achieve balance, normalization improves significantly. Although the raw accuracy fell slightly, the square justice and the lecturer increased. Especially, the MobileNet + Boosting appeared as a top artist (63.67%) on a balanced Fen dataset, demonstrating the power of light models in combination with the clothing classification for real-time and resource-wide scenarios.

Meanwhile, the base VIT, which weakened with unbalanced data, showed a significant improvement (61.24%) after the balance. This emphasizes the sensitivity to the Vit for data distribution and the ability to learn global functions when trained on equal class representatives. Adam Optimizer, LR = 0.001 (CNNS), 3×10^{-5} (VIT), and frequent training settings such as the use of certain hyperparameters, proved to be effective in architecture, validating their generality. Campaign methods (eg, XGBOOST) enhanced the discriminatory power of CNN functions further, and confirmed that the hybrid pipeline can improve the end-to-end model on a limited or unbalanced dataset. This study indicates that databalance, architectural efficiency, and hybrid are important for achieving strong turns. Strong performance is offered jointly to market light CNN, while balanced training enables Vit to appear as a competitive alternative.

6. Conclusion and Future Work

The study suggested a hybrid FER framework by combining CNN with a boosting algorithm and compared it to the transformer-based VIT model under fixed training conditions. The results emphasized the importance of dataset balance, which improved generalization and square disability. The MobileNet + Boosting obtained the best accuracy (63.67%) on a balanced fan, while the base Vit was particularly improved from 50.38%to 61.24%. The use of frequent hyperparameter tuning in architecture proved to be effective. Future work will expand

this pipeline to large FER datasets such as RAF-DB and AffectNet, and will detect multimodal and attention-based approaches to increase the strength of applications in the real world.

References

[1] S. B. Punuri *et al.*, "Efficient Net-XGBoost: An Implementation for Facial Emotion Recognition Using Transfer Learning," *Mathematics*, vol. 11, no. 3, Feb. 2023, doi: 10.3390/math11030776.

[2] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput Appl*, vol. 35, no. 32, pp. 23311–23328, Nov. 2023, doi: 10.1007/s00521-021-06012-8.

[3] R. Appasaheb Borgalli and S. Surve, "Deep Learning Framework for Facial Emotion Recognition using CNN Architectures," in *Proceedings of the International Conference on Electronics and Renewable Systems, ICEARS 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1777–1784. doi: 10.1109/ICEARS53579.2022.9751735.

[4] M. F. Alsharekh, "Facial Emotion Recognition in Verbal Communication Based on Deep Learning," *Sensors*, vol. 22, no. 16, Aug. 2022, doi: 10.3390/s22166105.

[5] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN)," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 2079–2089. doi: 10.1016/j.procs.2024.04.197.

[6] J. H. Chowdhury, Q. Liu, and S. Ramanna, "Simple Histogram Equalization Technique Improves Performance of VGG Models on Facial Emotion Recognition Datasets," *Algorithms*, vol. 17, no. 6, Jun. 2024, doi: 10.3390/a17060238.

[7] A. Vulpe-Grigorasi and O. Grigore, "Convolutional Neural Network Hyperparameters optimization for Facial Emotion Recognition," in *12th International Symposium on Advanced Topics in Electrical Engineering, ATEE 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021. doi: 10.1109/ATEE52255.2021.9425073.

[8] X. Xu, X. Wang, Z. Sun, and S. Wang, "Face recognition technology based on CNN, XGBoost, model fusion and its application for safety management in power system," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1755-1315/645/1/012054.

[9] S. Bobojanov, B. M. Kim, M. Arabboev, and S. Begmatov, "Comparative Analysis of Vision Transformer Models for Facial Emotion Recognition Using Augmented Balanced Datasets," *Applied Sciences (Switzerland)*, vol. 13, no. 22, Nov. 2023, doi: 10.3390/app132212271.

[10] F. Hernández-Luquin and H. J. Escalante, "Multi-branch deep radial basis function networks for facial emotion recognition," *Neural Comput Appl*, vol. 35, no. 25, pp. 18131–18145, Sep. 2023, doi: 10.1007/s00521-021-06420-w.

[11] Sarvakar, K. et al. 'Facial emotion recognition using convolutional neural networks', Materials Today: Proceedings, 80, pp. 3560–3564., 2023, doi:10.1016/j.matpr.2021.07.