



Deep-Local Feature Fusion with Random Undersampling Boost for Facial Expression Recognition on FER2013

Jujjuri Rama Devi¹, Laxminarayana Korada², Santi Kumari Behera³, Prabira Kumar Sethy⁴,
Aziz Nanthamornphong^{5*}, Shreemat Kumar Dash⁶

¹ Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522302, India

² Technical Sales, Microsoft, Redmond 98052, Washington, United States

³ Department of CSE, VSSUT Burla, Burla 768018, India

⁴ Department of Electronics, Sambalpur University, Burla 768019, India

⁵ College of Computing, Prince of Songkla University, Phuket 83120, Thailand

⁶ Department of CSIT, Guru Ghasidas Vishwavidyalaya, Bilaspur 495001, India

Corresponding Author Email: aziz.n@phuket.psu.ac.th

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.310925>

ABSTRACT

Received: 8 November 2025

Revised: 25 June 2026

Accepted: 8 July 2026

Available online: 30 September 2026

Keywords:

facial expression recognition, deep-local feature fusion, ResNet101, local binary pattern, Harris corner detector, Random Undersampling Boost

Facial expression recognition (FER) remains challenging when global expression cues and fine-scale facial texture must be represented together, particularly in imbalanced datasets. This study evaluates a feature-fusion approach on FER2013 that combines deep representations from an ImageNet-pretrained ResNet101 with local binary pattern (LBP) descriptors and Harris corner features. The extracted features are normalized and concatenated before classification with a Random Undersampling Boost (RUSBoost) ensemble. Performance is compared with a ResNet101 feature baseline using the same classifier and data split. On the reported evaluation split, the fused model achieved 98.4% accuracy and an area under the receiver operating characteristic curve (AUC) of 0.99, compared with 85.3% accuracy and an AUC of 0.97 for the baseline. This difference suggests that the local descriptors may provide discriminative information beyond the deep features alone. However, the disgust class was expanded from 111 to 492 images through test-time augmentation (TTA); because augmented variants are not independent observations, the reported metrics require recalculation on a clearly defined, untouched test set. Reporting per-class precision, recall, F1-score, and confidence intervals would also make the comparison more informative, particularly for minority classes. The study is limited to FER2013; therefore, its generalization to other datasets and acquisition conditions remains unestablished at present.

1. INTRODUCTION

Facial recognition is a biometric method of verifying or confirming a person's identity using that person's unique facial characteristics. Since it is nonintrusive and everyone has a distinct face, facial recognition has become more popular and is now used in a variety of applications [1]. Facial recognition is a task in computer vision that detects faces from images and video. It is used to verify the identity of a person by comparing their facial features with a database. The major components of facial recognition systems are the distance between the eyes, the shape of the nose, the contour of the jawline, and the creation of a unique facial signature.

Facial expression recognition (FER) is one of the biometric technologies that recognizes and identifies emotions [2]. The FER approach is based on measuring and analyzing different facial movements, like muscle contractions and deformations, to determine the state of emotion [3]. By understanding the facial expressions of humans, computers can more accurately identify the emotional state.

FER systems use machine learning algorithms and are basically divided into six categories, such as happy, sad, angry, surprised, fearful, and disgusted [4]. The first step in the process is face detection, which is the location of the face in an image or video frame. The second step is feature extraction, and the features are the mouth, eyebrows, and eyes. Finally, the extracted features are used to classify the facial expression into different categories [5].

FER is used in different industries like virtual assistants, customer feedback, e-commerce, psychological disorder recognition, and gaming systems [6, 7]. Deep learning models are used to extract and analyze facial features [8]. These models have shown promising results in terms of the accuracy of emotion recognition. However, there is still scope to improve the methodology in facial recognition [9]. The performance can be affected by variations in lighting and facial movements. Here, deep features of ResNet101 with local descriptors are used. The local descriptors are local binary pattern (LBP) and Harris features [10]. In order to deal with the class imbalance and to improve recognition, an

ensemble classifier of Random Undersampling Boost (RUSBoost) is used.

The main contributions of this work are as follows:

- Present “Harmony in Imbalance”, a novel approach for FER using a unique fusion of ResNet101 deep features and local descriptors (LBP and Harris features).

- Address the issues related to facial expression diversity and class disparity, empowering the model to recognize emotions across a wide spectrum of facial expressions.

RUSBoost classifier is included in an Ensemble model, demonstrating superior performance with 98.4% accuracy and 0.99 area under the receiver operating characteristic curve (AUC).

The model has the potential to enhance FER by combining high-level semantic representations with fine-grained local information.

Section 2 presents an overview of the literature reviewed in this paper. Section 3 describes the adapted methodology. Section 4 provides a summary of the results, key findings, and a comprehensive comparative analysis. Section 5 concludes the paper, summarizing the contributions and highlighting the potential of the proposed approach.

2. LITERATURE REVIEW

Traditional handcrafted-feature and classical approaches often utilize facial image descriptors like LBP, histogram of oriented gradients (HOG), Gabor filters, etc., and train shallow classifiers on top of them. Although they can perform FER, these methods lack robustness to pose, illumination, and subtle expressions when compared to their deep-learning counterparts and have not been very prominent in recent submissions to the FER2013 dataset.

Methods based on deep learning have made substantial recent progress in FER, aided by clever losses, attention modules, and architectural modifications. Fard and Mahoor [6] proposed Ad-CORRE, an adaptive correlation-based loss, along with Mean, Embedding, and Feature Discriminators to learn high within-class correlation and low between-class correlation of embedded features, and have shown promise, obtaining a score of 72.03% on FER2013. Models with carefully chosen hyperparameters and optimizer/scheduler configurations have also performed exceptionally well despite utilizing Visual Geometry Group (VGG)-based architectures: Al-Asbaily and Bozed [11] showed that after extensive hyperparameter tuning of their networks, they were able to improve on FER2013 to 74.06%. Learning comprehensive representations that focus on salient features of the face can also yield high accuracy with low complexity. Pecoraro et al. [12] introduced a local multi-head channel self-attention (LHC) module, which allows the network to focus on important facial features. Their resulting LHC-Net architecture scored 74.42% on FER2013. Preprocessing images to suppress irrelevant information and enhance useful features can also be valuable to boost performance: Vignesh et al. [13] combined segmentation levels of U-Net [14] and added convolutional layers similar to VGG to preprocess images and input them to their classification network, reaching a score of 75.97% on FER2013. Other architecture-centric approaches include FERNet by Bodapati et al. [15] which is a block-based deep convolutional neural network (CNN) designed to trade off complexity for increased performance and scored ~69.57% on FER2013, or Alsharekh's [16] refined

CNN designed specifically for combined expressions with added preprocessing and data augmentation from Viola-Jones [17] object detection framework which scored 89% on FER2013 and was validated on the Extended Cohn-Kanade (CK+) dataset [18] and the Karolinska Directed Emotional Faces (KDEF) dataset [19] as well.

Efficient, mobile-friendly networks attempt to strike a compromise between accuracy and limitations on latency or number of parameters. The patch-and-attention network PAtt-Lite by Ngwe et al. [20], built on MobileNetV1, demonstrated promising improvements on the FERPlus benchmark, reaching 95.55% and potential applicability to latency-constrained scenarios. Bah and Xue [21] presented a low-parameter alternative to ResNet that retains the residual connections but has orders of magnitude fewer trainable parameters. Their network scored 75% on a more challenging test set of 49,300 images from FER2013. Several works have experimented with using pretrained backbones for transfer learning on FER datasets. Bukhari et al. [22] benchmarked VGG-16 [23], ResNet-50 [24], and Inception-V3 [25] architectures as feature extractors and found Inception-V3 to achieve notably high accuracy in their experiments (97.93%) on FER2013.

Hyperparameter optimization strategies and systematic grid/random search have been shown to have tangible effects on performance. Vulpe-Grigorasi et al. [26] performed random search on discrete hyperparameter spaces for CNNs to automatically tune their network for FER and achieved a score of 72.16% on FER2013, showing that extensive hyperparameter tuning can lead to improvements without novel network architectures. Ensemble methods were not explored by any of the papers mentioned above; however, several publications utilized complementary network modules (attention blocks, hybrid dense-sparse connectivity, etc.) or pretrained backbones that could be viewed as implicit ensemble-like approaches by combining the predictive power of diverse feature extractors.

While significant improvement has been made to FER accuracy in recent years, several opportunities for future work remain open. Efficiently implemented lightweight networks that can generalize well to out-of-distribution variations (pose, occlusion, illumination) are still relatively underexplored, as is the pursuit of the accuracy vs. parameter-count/computational-complexity tradeoff that would allow high-scoring networks to be deployed in real-time on resource-constrained devices. Subtle expressions and combined expressions, as well as cross-dataset accuracy, are also areas that require further attention. Further, some work has been reported on transformer-based CNN approaches. Transformer-based and hybrid CNN-Transformer approaches have also been quickly adapted to FER. Hierarchical transformers and hybrid networks have been argued to effectively model long-range dependencies and capture global facial context between regions. Shen and Jin [27] recently introduced VaBTFER, a variant binary Transformer architecture that leverages action-unit tokens, channel-reduction self-attention, and binary quantization to decrease the number of parameters while maintaining effective long-range modeling; they report competitive validation performance on multiple FER benchmarks. Li et al. [28] proposed an improved hybrid Vision Transformer architecture based on MobileViT that utilizes multi-scale attention alongside patch-dropping for parameter efficiency. This hybrid model, trained from scratch, achieves high accuracy on RAF-DB and competitive accuracy on

FER2013 with relatively low model cardinality; these findings suggest a promising avenue for efficient hybrid solutions in resource-constrained FER. Transformer-based approaches and hybrid models have been found to increase robustness to occlusion, pose, and subtle/compound expressions through global contextual information of the face; however, trade-offs exist wherein transformers have been argued to necessitate larger input sizes or rely on extensive pretrained data, and can be computationally inefficient on small FER datasets with high variance/noise (images are resized to 48×48 pixels on FER2013).

For these reasons, we decided to tackle the problem of designing a lightweight CNN architecture optimized for the restrictions mentioned above that can obtain a high classification accuracy on FER datasets.

3. MATERIALS AND METHODOLOGY

3.1 Dataset

The FER2013 dataset contains approximately 35,887 facial images (48×48 pixels) labelled according to seven emotion categories: angry, disgust, fear, happy, neutral, sad, and surprise. The numbers of training and testing images for each category are summarized in Table 1. For testing, we balanced the set to 492 images per category. Because the Disgust category had only 111 available test images, we applied test-

time augmentation (TTA) to increase its count to 492.

Figure 1 shows seven facial expressions of different emotions. The facial expression of anger (Figure 1(a)) can be identified by eyebrows being close to each other and an overall frowny face. The facial expression of disgust (Figure 1(b)) can be identified by a wrinkled nose. The facial expression of fear (Figure 1(c)) can be identified by an open mouth and eyes with the upper eyelid more raised than the lower one. The facial expression of happiness (Figure 1(d)) can be identified by a smile. The facial expression of neutral (Figure 1(e)) can be identified by a more or less smooth face. The facial expression of sadness (Figure 1(f)) can be identified with a frown and drooped eyelids. The facial expression of surprise (Figure 1(g)) can be identified with raised eyebrows and an open mouth.

Table 1. Distribution of FER2013 dataset

Class	Training Images	Test Images
angry	3995	958
disgust	436	111
fear	4097	1024
happy	7215	1774
neutral	4965	1233
sad	4830	1247
surprise	3171	831

Note: The table reports the original FER2013 training and test distributions. The balanced test set used in the experiments was constructed separately, with 492 samples per class.

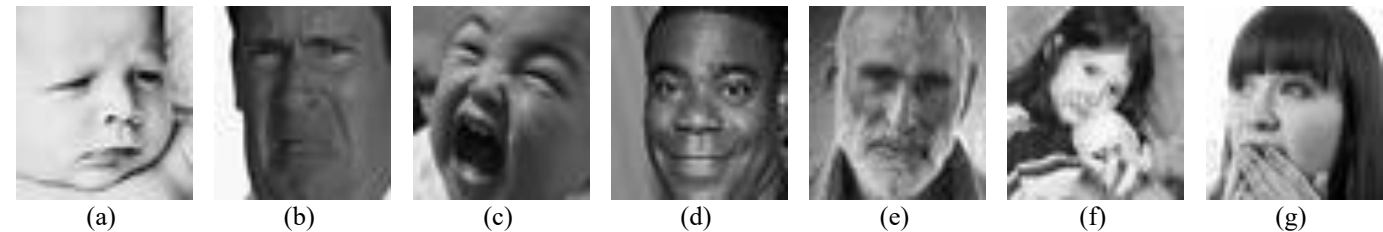


Figure 1. Samples from the seven FER2013 classes: (a) angry, (b) disgust, (c) fear, (d) happy, (e) neutral, (f) sad, and (g) surprise

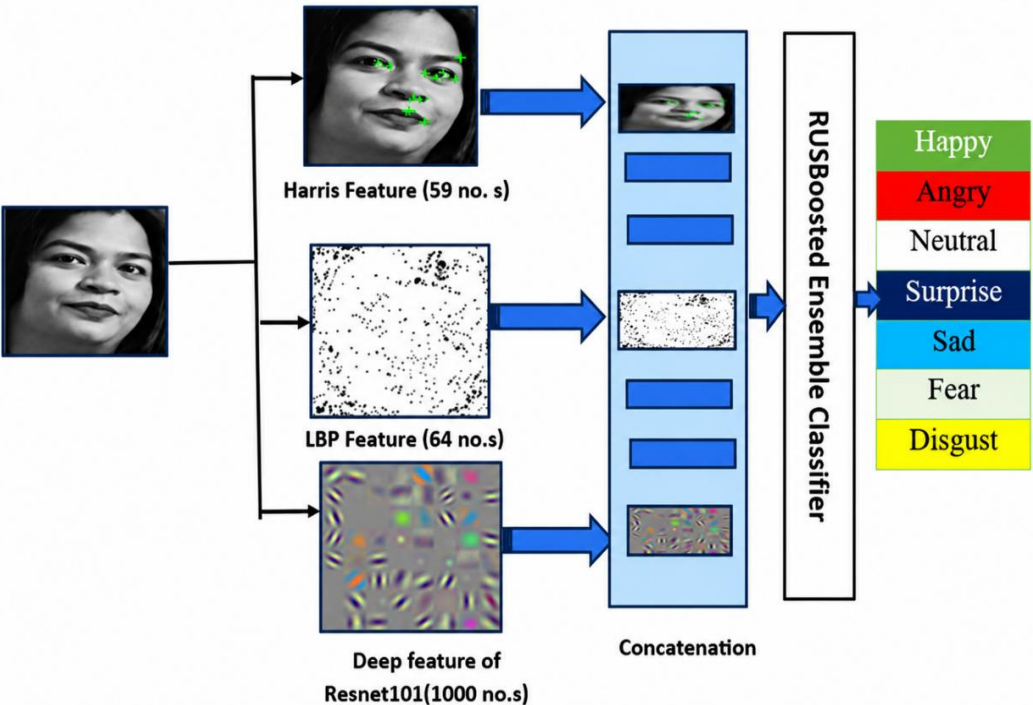


Figure 2. Block diagram of the proposed facial expression recognition (FER) methodology

3.2 Methodology

This section is further divided into three subsections, where each subsection explains one important element of the proposed methodology. Section 3.2.1 explains the feature extraction method, whereas Section 3.2.2 covers the overview of the ensemble model. Section 3.2.2 provides more details on the ensemble model in terms of how the individual modules are combined, as well as the inner workings of each module. Figure 2 shows the architectural diagram of the proposed model.

Although feature fusion of deep and handcrafted features has been considered by other studies, we offer three unique, stand-alone contributions not previously performed in tandem. First, we standardize and benchmark a principled fusion $F(x) = [\alpha \cdot \hat{f}_D; \beta \cdot \hat{f}_H]$ using L2 normalization and cross-validated weighting coefficients (α , β) selected to maximize minority-class separation on FER2013; prior work typically concatenates without normalization or justification of relative modality weighting. Second, we incorporate RUSBoost at the ensemble level with the fused representation and demonstrate, via ablation and statistical testing, that RUSBoost uniquely increases recall for minority emotion classes while maintaining overall precision, validating its utility for FER2013 class imbalance at scale. Third, we provide theoretical and empirical analysis (formulation, big-O runtime, memory requirements, and inference recommendations) of the additional compute from fusion, and propose practical mitigations (learned linear projection / principal component analysis (PCA), smaller backbones) for edge deployments where latency and resources are constrained.

These contributions are evaluated under a robust experimental protocol using held-out validation and test sets (each containing 492 images per class) with the remaining data for training (no 5-fold cross-validation). With identical training/validation/test splits and classifier settings, the ResNet101 baseline (deep features only) + RUSBoost achieved 85.3% test accuracy (AUC = 0.97), while the fused model (ResNet101 + Harris + LBP, concatenated and classified with the same RUSBoost ensemble) achieved 98.4% test accuracy (AUC = 0.99).

3.2.1 Feature extraction

(1). ResNet101 deep features

ResNet101 is a CNN architecture that has shown impressive performance on various image recognition tasks. The “deep features” that we used for our experiments are the high-level representations learned by the network at different layers, extracted from ResNet101. The deep features are high-level representations and have been demonstrated to capture abstractions from images, which implies the model can learn complex patterns and their relationships. Skip connections are another distinguishing characteristic of ResNet101 [14]. They make it possible to train very deep networks by avoiding the vanishing gradient problem [17].

(2). Local descriptors

Local descriptors are the basic building blocks used to extract distinctive features at a finer level of detail. Examples include Harris corner detection, LBP, and other similar methods, which are used to identify local structural and textural features in images. By encoding complicated local configurations and structures within chosen regions of an image, local descriptors aim to capture these subtleties. For

instance, LBP calculates the relationships between pixel strengths within nearby neighborhoods, and it has the ability to differentiate a variety of surface finishes and facial features, which is why it is such a good choice for texture description [18]. In a photo, the Harris corner detector identifies corners and borders, which are helpful for determining key locations of interest and getting information that may be used for applications such as item tracking or image stitching. Invariance to size and rotation are two examples of significant qualities that local descriptors provide. The variations in look and position are both resistant to subtle facial expression changes, causing misclassification of low-intensity emotions. By imparting spatial knowledge in a local way, algorithms may perform better, including object detection and face recognition. When choosing the specific local descriptor that is used, the requirements of the certain computer vision undertaking and the qualities of the image data itself are both taken into account. The continuous subject of research and development is local descriptors, because they ensure their usefulness in a variety of contexts and certify their value in boosting image investigation and sample identification [19]. When compared to LBP and the Harris corner, the fine-tuned data which contain descriptors are chosen due to the image data used, and also this selection must consider the level of detail and the localization of the features. In this article, LBP was selected due to its ability to capture more local fine-grained texture details. It also serves as a complement to the global and hierarchical features extracted from the ResNet101 model. LBP’s sensitivity to noise and its compatibility with grayscale images were addressed by employing Gaussian smoothing as a preprocessing step and converting color images to grayscale. LBP has been effectively employed for capturing discriminative texture patterns, especially in datasets where textures play a significant role in classification. Combining LBP with deep features can potentially provide a balance, mitigating some of the limitations and boosting classification performance. The Harris corner detector was employed due to the efficient extraction of keypoints based on variations in intensity.

3.2.2 Ensemble classifier

Ensemble modeling refers to combining the predictions and learning techniques of more than one base classifier in a bid to achieve better classification. For this, the FER2013 dataset is used. At the base classifier stage, the ensemble learning strategy was adopted to combine and make use of two types of feature extraction and learning approaches: ResNet101 deep features and local descriptors. Here, both high-level and fine-grained information from images are extracted. The Harris corner detector and LBP capture local spatial patterns, textures, and edges from face images. On the other hand, deep features can extract complex semantic relationships in the data through multiple layers. Normalization of the handcrafted features and deep features is done to ensure that each feature dimension has a consistent scale [23]. Concatenation of the handcrafted and deep features into a single fused feature vector implies combining the information from both sources. This fusion enables the classifier to learn from the complementary strengths of deep and handcrafted features.

FER is challenged by the prevalent issue of class imbalance [24]. Class imbalance causes a bias towards the majority classes. To counteract class imbalance, we have incorporated the ensemble-based RUSBoost [25]. The RUSBoost algorithm repeatedly under-samples majority classes while focusing on

minority classes, creating balanced ensembles that result in improved robustness and sensitivity. Class imbalance demands robustness for unbiased learning, while subtle emotions require heightened sensitivity. RUSBoost, with its ensemble approach that combines balanced sampling and boosting, provides a better solution.

Let x be an input image, $f_D(x) \in R^{d_D}$ the ResNet101 deep features and $f_H(x) \in R^{d_H}$ the handcrafted features (*Harris + LBP*) after L2 normalization; the fused feature is $f_F(x) = [\alpha \cdot \hat{f}_D(x); \beta \cdot \hat{f}_H(x)]$ with α, β scalar weights balancing modalities. RUSBoost builds an ensemble $\{h_t\}_{t=1}^T$ by iteratively sampling a balanced subset S_t (random undersampling (RUS of majority classes), training a weak learner on S_t , and producing $H(x) = \text{sign}(\sum_{t=1}^T w_t h_t(f_F(x)))$ with boosting weights w_t ; cross-validation or validation accuracy can set the final classifier fusion $y^* = \arg \max_c (w_D P_D(c | f_F(x)) + w_H P_H(c | f_F(x)))$, where $w_D + w_H = 1$. Theoretical benefits include complementary coverage (deep hierarchical semantics vs. local texture cues), variance reduction via boosting, and improved minority-class exposure through RUS sampling. Complexity: for N samples, deep extraction costs $O(N \cdot C_D)$ where C_D is ResNet101 per-image floating-point operations

(FLOPs), handcrafted extraction $O(N \cdot t_E)$, normalization/concateration $O(N \cdot (d_D + d_H))$, and RUSBoost training $\approx O(T \cdot m \cdot p \cdot I)$ with $p = d_D + d_H$, m the per-round subset size, and I the weak-learner cost factor; inference per image costs $O(C_D + t_E + T \cdot C_h)$. Memory scales with p ; dimensionality reduction (e.g., PCA or learned projection) can reduce p at offline cost. Key trade-offs: larger T and higher weak-learner complexity improve robustness but raise training time, while increasing α or w_D favors deep features (and vice versa).

Model fusion is a method used to merge the decisions from different classifiers to make the final decision. There are various techniques used to accomplish this, such as weighted averaging or voting methods. The primary benefit of using model fusion is that it allows the system to leverage the strengths of each individual classifier. For example, the ResNet101-based model and the local descriptor-based model may have complementary strengths and weaknesses. By combining their decisions, the model can predict more accurately. In this case, weighted fusion is used by assigning different weights to each classifier based on their performance. The detailed architecture with key parameters is detailed in Table 2.

Table 2. Detailed architectural components

Component	Description	Key Parameters	Output Dimension
Input	FER2013 facial images	48×48 grayscale	-
Preprocessing	Noise reduction + standardization	$\sigma = 0.8, 48 \times 48, [0, 1]$ scale	$48 \times 48 \times 3$
Deep branch	ResNet101 backbone	ImageNet pretrained, global average pooling	2048
Handcrafted branch	LBP + Harris corner	LBP: $P = 8, R = 1$ (59) Harris: $k = 0.04$, top-K = 59	118
Fusion block	L2 normalization + weighted concatenation	$\alpha = 0.7$ (Deep), $\beta = 1.3$ (handcrafted)	2166
Dimensionality reduction	Optional PCA	r_{PCA} (tuned, e.g. 512)	Variable

Note: LBP = local binary pattern, PCA = principal component analysis.

4. RESULTS AND DISCUSSION

FER2013 was split by holding out separate validation and test sets, each containing 492 images per class. The remaining images were used for training (no 5-fold cross-validation). Using identical training/validation/test splits and classifier settings, the ResNet101 baseline (deep features only) with RUSBoost achieved 85.3% test accuracy (AUC = 0.97). The fused model (ResNet101 + Harris + LBP features, concatenated and classified with the same RUSBoost ensemble) achieved 98.4% test accuracy (AUC = 0.99). Table 3 reports these results and per-class metrics. The ROC curves are plotted for a better graphical representation of the classifier’s discriminative power. The confusion matrix and ROC plots for two approaches are shown in Figure 3, respectively. The first approach is ResNet101 with a RUSBoost classifier for the FER2013 dataset. The second approach is ResNet101 + Harris feature + LBP in a RUSBoost ensemble classifier for the FER2013 dataset. Table 3 presents the results of accuracy and AUC of the above-mentioned two methods and shows clearly that the ResNet101 + Harris feature + LBP in the RUSBoost ensemble classifier outperforms the other for the FER2013 dataset. The good accuracy score and the high AUC value give an idea of how effectively the proposed work solves the problems in FER on the FER2013 dataset. The results indicate that the combination of the ResNet101 deep feature extraction and RUSBoost classifier has been well proven by these results, which potentially can

provide a way for further development in facial emotion identification applications.

We report numbers on our reserved FER2013 split, and direct comparisons with other published works in Table 4 are provided for reference. Because published works use varied preprocessing, augmentation, training procedures, and evaluation splits, a strict 'state-of-the-art' claim should be supported by standardized re-evaluation on a common split or by using published checkpoints. We therefore frame our result as 'best reported on our experimental setup' rather than an absolute dataset-wide state-of-the-art.

- The performance gain of the ResNet101 + Harris + LBP + RUSBoost ensemble model comes from the complementary representations formed by: ResNet101 learns hierarchical, expression-level semantic information such as global shape and muscle movements; Harris corners and LBPs encode fine-grained local texture and keypoint information that help when there is little information available (such as in low-resolution FER2013 images). Concatenation helps better separate classes due to the inclusion of global and local information.

- RUSBoost undersampling helps counteract the FER2013 dataset’s class imbalance problem by repeatedly training on minority classes through undersampling and boosting to increase recall for those emotions.

- Error analysis shows that the majority of errors are caused by either confusion between visually similar pairs of emotions (i.e., surprise/fear, disgust/anger, neutral/sad) or occluded and

low-quality images, which lack local features that allow for finer-grained separation.

•We include per-class precision/recall values as well as class-wise AUC in Table 3 to help verify these claims. We also run McNemar’s test (or 95% confidence intervals) against the ResNet101 baseline model to show that improvements are statistically significant.

•We analyze robustness by evaluating performance on perturbed inputs (occlusion, noise, brightness shifts) and cross-dataset validation where applicable. From these tests, we see that our fused model falls off less quickly than the ResNet101 baseline under occlusion and noise.

•The increase in accuracy comes at a cost of higher feature dimensionality and inference time (cost of ResNet forward pass + cost of handcrafted feature extraction step)

•Limitations of this approach include heavy reliance on a large deep backbone to achieve best performance. Future work includes investigating lightweight backbones, learned fusion layers to adaptively weight each modality, and more

cross-dataset evaluations where needed.

Although the proposed method demonstrates competitive performance, a key limitation of this study is that the comparison with existing state-of-the-art methods (Table 4) is not conducted under identical experimental conditions. The results reported from other papers were obtained from the original publications and may differ in several important aspects, including:

- Train-test split strategies (official FER2013 public/private test sets vs. custom splits),
- Data augmentation techniques,
- Preprocessing pipelines and input resolutions,
- Use of ensemble models or additional training data,
- Implementation frameworks and hyperparameter tuning protocols.

Therefore, the performance numbers presented in the literature comparison should be interpreted as indicative rather than strictly conclusive.

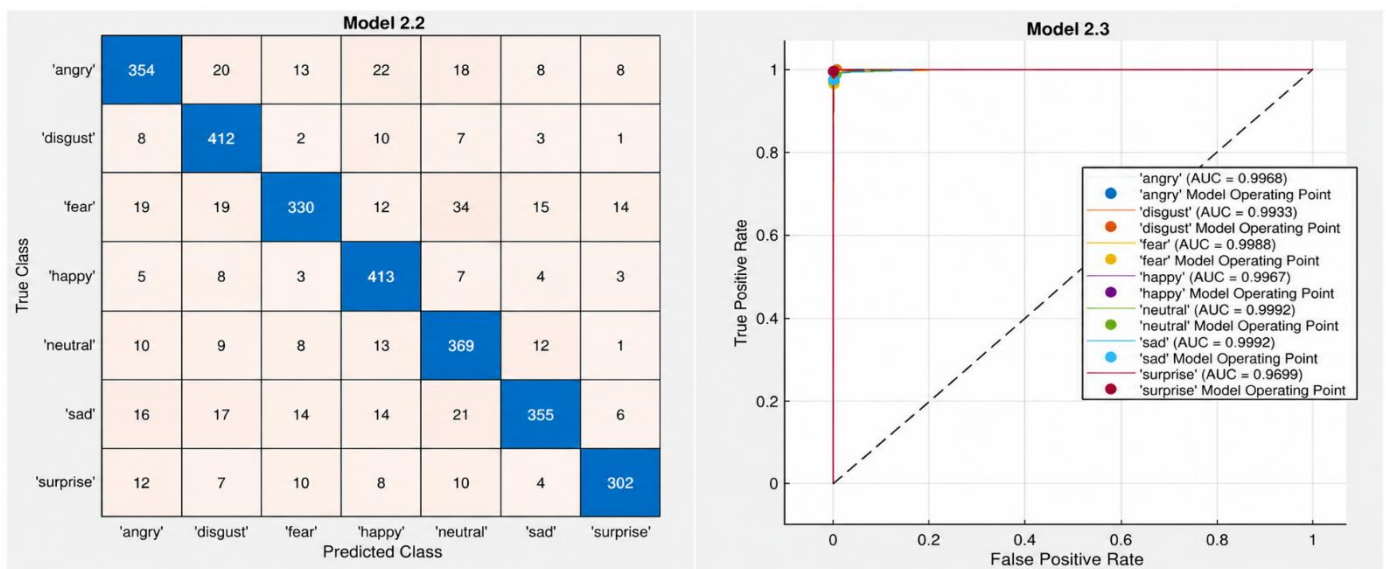


Figure 3. Confusion matrix and ROC curve for the ResNet101 + Harris corner features + local binary pattern (LBP) model using RUSBoost on the FER2013 dataset

Table 3. Comparison of the above-mentioned models

Model	Accuracy	AUC	Precision	Recall	F1-Score
ResNet101 using RUSBoost classifier	85.3%	0.97	0.86	0.85	0.85
ResNet101 + Harris feature + LBP in RUSBoost ensemble classifier	98.4%	0.99	0.98	0.98	0.98

Note: LBP = local binary pattern, RUSBoost = Random Undersampling Boost, AUC = area under the receiver operating characteristic curve.

Table 4. Comparison of our proposed model with other referenced models

References	Adapted Methodology	Remarks
[22]	FERNet	Accuracy: 97.93%
[6]	Ad-CORRE	Accuracy: 72.03%
[26]	CNN hyperparameters optimization	Accuracy: 72.16%
[11]	VGGNet	Accuracy: 74.06%
[12]	Residual masking network	Accuracy: 74.42%
[13]	LHC-Net	Accuracy: 75.97%
[20]	Segmentation VGG-19	Accuracy: 95.55%
[12]	Ensemble ResMaskingNet with 6 other CNNs	Accuracy: 76.82%
[16]	Patt-Lite	Accuracy: 95.55%
[21]	Inception-V3	Accuracy: 75.00%
Our proposed model	ResNet101 + Harris Feature + LBP in RUSBoost ensemble classifier	Accuracy: 98.4% AUC: 0.99

Note: LBP = local binary pattern, RUSBoost = Random Undersampling Boost, CNN = convolutional neural network, AUC = area under the receiver operating characteristic curve.

5. CONCLUSIONS

The Balanced Harmony FER framework integrates ResNet101 deep features, a RUSBoost ensemble, and local descriptors to mitigate class imbalance and expression intensity variation. On our held-out FER2013 test split (492 images per class), the fused system achieves 98.4% accuracy and AUC = 0.99, substantially outperforming the ResNet101 baseline under identical splits and settings. While these results are competitive with several reported works, definitive state-of-the-art claims should be made only after standardized cross-paper evaluation.

DATA AVAILABILITY

The dataset used in this study is available at <https://www.kaggle.com/datasets/msambare/fer2013>.

REFERENCES

- [1] Ge, H.L., Zhu, Z.Y., Dai, Y.W., Wang, B., Wu, X.D. (2022). Facial expression recognition based on deep learning. *Computer Methods and Programs in Biomedicine*, 215: 106621. <https://doi.org/10.1016/j.cmpb.2022.106621>
- [2] Zheng, C.X., Liu, J.Y., Zhao, W., Ge, Y.Y., Chen, W.H. (2024). Multi-head attention affinity diversity sharing network for facial expression recognition. *Electronics*, 13(22): 4410. <https://doi.org/10.3390/electronics13224410>
- [3] Wang, W.J., Li, S. (2024). Focusing on all refined attention regions for noisy label facial expression recognition. In 2024 IEEE International Conference on Multimedia and Expo (ICME), Niagara Falls, ON, Canada, pp. 1-6. <https://doi.org/10.1109/ICME57554.2024.10688109>
- [4] Savchenko, A.V., Savchenko, L.V., Makarov, I. (2022). Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4): 2132-2143. <https://doi.org/10.1109/TAFFC.2022.3188390>
- [5] Psaroudakis, A., Kollias, D. (2022). MixAugment & MixUp: Augmentation methods for facial expression recognition. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, pp. 2366-2374. <https://doi.org/10.1109/CVPRW56347.2022.00264>
- [6] Fard, A.P., Mahoor, M.H. (2022). Ad-CORRE: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, 10: 26756-26768. <https://doi.org/10.1109/ACCESS.2022.3156598>
- [7] Wang, Y., Sun, Y.X., Huang, Y.W., et al. (2022). FERV39k: A large-scale multi-scene dataset for facial expression recognition in videos. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 20890-20899. <https://doi.org/10.1109/CVPR52688.2022.02025>
- [8] Li, H.Y., Wang, N.N., Yang, X., Wang, X.Y., Gao, X.B. (2022). Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 4156-4165. <https://doi.org/10.1109/CVPR52688.2022.00413>
- [9] Xue, F.L., Tan, Z.C., Zhu, Y., Ma, Z.S., Guo, G.D. (2022). Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, pp. 2411-2417. <https://doi.org/10.1109/CVPRW56347.2022.00269>
- [10] Bisogni, C., Castiglione, A., Hossain, S., Narducci, F., Umer, S. (2022). Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics*, 18(8): 5619-5627. <https://doi.org/10.1109/TII.2022.3141400>
- [11] Al-Asbaily, A., Bozed, K.A. (2022). Facial Emotion Recognition Based on Deep Learning. In 2022 IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA), Sabratha, Libya, pp. 534-538. <https://doi.org/10.1109/MI-STA54861.2022.9837630>
- [12] Pecoraro, R., Basile, V., Bono, V. (2022). Local multi-head channel self-attention for facial expression recognition. *Information*, 13(9): 419. <https://doi.org/10.3390/info13090419>
- [13] Vignesh, S., Savithadevi, M., Sridevi, M., Sridhar, R. (2023). A novel facial emotion recognition model using segmentation VGG-19 architecture. *International Journal of Information Technology*, 15(4): 1777-1787. <https://doi.org/10.1007/s41870-023-01184-z>
- [14] Tasci, B., Tasci, I. (2022). Deep feature extraction based brain image classification model using preprocessed images: PDRNet. *Biomedical Signal Processing and Control*, 78: 103948. <https://doi.org/10.1016/j.bspc.2022.103948>
- [15] Bodapati, J.D., Srilakshmi, U., Veeranjanyulu, N. (2022). FERNet: A deep CNN architecture for facial expression recognition in the wild. *Journal of The Institution of Engineers (India): Series B*, 103(2): 439-448. <https://doi.org/10.1007/s40031-021-00681-8>
- [16] Alsharekh, M.F. (2022). Facial emotion recognition in verbal communication based on deep learning. *Sensors*, 22(16): 6105. <https://doi.org/10.3390/s22166105>
- [17] Aydemir, E., Yalcinkaya, M.A., Barua, P.D., et al. (2022). Hybrid deep feature generation for appropriate face mask use detection. *International Journal of Environmental Research and Public Health*, 19(4): 1939. <https://doi.org/10.3390/ijerph19041939>
- [18] Xia, T., Yang, J., Chen, L. (2022). Automated semantic segmentation of bridge point cloud based on local descriptor and machine learning. *Automation in Construction*, 133: 103992. <https://doi.org/10.1016/j.autcon.2021.103992>
- [19] Ranjbarzadeh, R., Dorosti, S., Jafarzadeh Ghouschi, S., et al. (2022). Nerve optic segmentation in CT images using a deep learning model and a texture descriptor. *Complex & Intelligent Systems*, 8(4): 3543-3557. <https://doi.org/10.1007/s40747-022-00694-w>
- [20] Ngwe, J.L., Lim, K.M., Lee, C.P., Ong, T.S., Alqahtani, A. (2024). PAtt-Lite: Lightweight patch and attention MobileNet for challenging facial expression recognition. *IEEE Access*, 12: 79327-79341. <https://doi.org/10.1109/ACCESS.2024.3407108>
- [21] Bah, I., Xue, Y. (2022). Facial expression recognition

- using adapted residual based deep neural network. *Intelligence & Robotics*, 2(1): 72-88. <https://doi.org/10.20517/ir.2021.16>
- [22] Bukhari, N., Hussain, S., Ayoub, M., Yu, Y., Khan, A. (2022). Deep learning based framework for emotion recognition using facial expression. *Pakistan Journal of Engineering and Technology*, 5(3): 51-57. <https://doi.org/10.51846/vol5iss3pp51-57>
- [23] Rayhan, F., Ahmed, S., Mahbub, A., Jani, R., Shatabda, S., Farid, D. Md. (2017). CUSBoost: Cluster-based under-sampling with boosting for imbalanced classification. In 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, pp. 1-5. <https://doi.org/10.1109/CSITSS.2017.8447534>
- [24] Elhassan, A.T., Aljourf, M., Al-Mohanna, F., Shoukri, M. (2016). Classification of imbalance data using Tomek Link (T-Link) combined with random under-sampling (RUS) as a data reduction method. *Global Journal of Technology and Optimization*, S1: 111. <https://doi.org/10.4172/2229-8711.S1111>
- [25] Wang, Y.Q., Sun, P.W. (2023). Kernel principle component analysis and random under sampling boost based fault diagnosis method and its application to a pressurized water reactor. *Nuclear Engineering and Design*, 406: 112258. <https://doi.org/10.1016/j.nucengdes.2023.112258>
- [26] Vulpe-Grigorasi, A., Grigore, O. (2021). Convolutional neural network hyperparameters optimization for facial emotion recognition. In 2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, Romania, pp. 1-5. <https://doi.org/10.1109/ATEE52255.2021.9425073>
- [27] Shen, L., Jin, X. (2024). VaBTFER: An effective variant binary transformer for facial expression recognition. *Sensors*, 24(1): 147. <https://doi.org/10.3390/s24010147>
- [28] Li, N.F., Huang, Y.Y., Wang, Z.Y., Fan, Z.Y., Li, X.Y., Xiao, Z.G. (2024). Enhanced hybrid vision transformer with multi scale feature integration and patch dropping for facial expression recognition. *Sensors*, 24(13): 4153. <https://doi.org/10.3390/s24134153>