Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 30 May 2026
Revised 20 June 2026
Accepted 18 July 2026

Natural Resources for Human Health 2026, Vol. 6 (9s): 868–878

KEYWORDS:
convolutional neural network, CK+ dataset, deep learning, EmotionNet, facial emotion recognition, real time video analysis, transfer learning, VGG16.

Comparative Performance Analysis of a Custom CNN, VGG16, and the Proposed Emotion Net Model for Facial Emotion Recognition

Amit B. Rehapade[1], Dr. Prafulla E. Ajmire[2], Nikhil V. Khandar[3], Shubhangi Amit Gahane[4]

[1]Research Scholar, Department of Computer Science and Engineering, S.G.B Amravati University, Amravati - India
[2]Supervisor, Research Centre, Department of Computer Science and Engineering, S.G.B Amravati University, Amravati - India
[3]Assistant Professor, Dr. Ambedkar Institute of Management Studies & Research, Nagpur, Maharashtra – India
[4]Assistant Professor, Department of Computer Science, Santaji Mahavidyalaya, Nagpur- India

ABSTRACT: Facial emotion recognition (FER) plays a pivotal role in affective computing, healthcare monitoring, adaptive education, surveillance, and human computer interaction. This study benchmarks three deep learning solutions for seven class FER on the Extended Cohn Kanade (CK+) corpus: a compact custom convolutional neural network (CNN), a VGG16 model adapted through transfer learning, and a newly proposed architecture, EmotionNet. All three models are trained and evaluated under an identical preprocessing and augmentation protocol so that observed performance differences can be attributed to network design rather than to inconsistencies in data handling. The custom CNN is inexpensive to run, whereas VGG16 leverages pretrained visual features at the cost of considerably higher memory and computational demand. EmotionNet is built from stacked convolutional blocks with batch normalization, rectified linear unit (ReLU) activation, max pooling, and dropout, followed by fully connected classification layers, and it is embedded within an end to end real time video inference pipeline. On the CK+ benchmark, the three models reach classification accuracies of 92.84%, 96.73%, and 98.86%, respectively, with EmotionNet additionally producing 98.54% precision, 98.31% recall, and a 98.42% F1 score using 42.70 million parameters and 7.82 GFLOPs of computation. These findings indicate that EmotionNet delivers the strongest recognition accuracy while maintaining a more favorable computation to accuracy trade off than VGG16, whereas the custom CNN remains the most suitable choice when hardware resources are severely constrained.

I. INTRODUCTION

Human facial expressions convey a person's emotional state through coordinated changes in the eyes, eyebrows, mouth, and surrounding musculature. Automated facial emotion recognition (FER) translates these visual cues into discrete emotion categories and, in doing so, underpins a wide range of emotion aware interfaces and decision support systems. Although deep learning has substantially advanced FER, the task remains challenging in practice because factors such as expression intensity, illumination, head pose, partial occlusion, and inter subject variability can alter how the same emotion is visually expressed [1], [2].

Traditional FER pipelines separate handcrafted feature extraction from a downstream classifier, whereas convolutional neural networks (CNNs) learn hierarchical, task relevant representations directly from raw pixel data. Transfer learning extends this capability by reusing features learned from large scale image corpora, thereby improving data efficiency on smaller, domain specific datasets. VGG16 is a commonly adopted transfer learning backbone because its stacked 3×3 convolutional filters yield strong spatial feature representations [3]; however, its large parameter count and computational footprint restrict its use in real time or resource constrained deployments. Lightweight, purpose built CNNs address this efficiency gap but can lose discriminative power when expressions are subtle or visually similar to one another.

This paper positions a lightweight custom CNN and a VGG16 transfer learning model as baselines and introduces EmotionNet, a purpose built architecture designed to reconcile recognition accuracy, generalization, and real time feasibility. All three models are trained and tested under a single, controlled CK+ evaluation protocol so that the comparison isolates the effect of network design. The principal contributions of this work are as follows:

• A unified preprocessing and augmentation pipeline that enables a controlled, three way model comparison;

• The design of EmotionNet, which couples deep hierarchical feature extraction with regularization and stable optimization;

• A per emotion evaluation reported in terms of accuracy, precision, recall, and F1 score; and

• A joint assessment of recognition quality, parameter count, and floating point operations (FLOPs), memory footprint, and overall deployment suitability.

II. RELATED WORK

Early approaches to FER relied on handcrafted appearance descriptors and geometric facial measurements combined with conventional classifiers. The introduction of deep CNNs removed much of the dependence on manual feature engineering by learning low and high level facial patterns jointly from data. AlexNet first demonstrated the power of large scale convolutional learning for visual recognition [4], and subsequent work on very deep networks with small convolutional kernels, such as VGG, showed that increasing network depth further improves the quality of learned visual representations [3]. Training stability and generalization were later improved through batch normalization [5], dropout regularization [6], and adaptive gradient based optimizers such as Adam [7].

More recent FER research has explored deep residual networks, mobile oriented architectures, attention mechanisms, and transfer learning. Residual connections enable the training of substantially deeper networks without degradation [8], while MobileNet reduces computational cost through depthwise separable convolutions, making it attractive for embedded inference [9]. Commonly used FER benchmarks include FER2013, AffectNet, FERPlus, and CK+ [12]–[14]; among these, CK+ remains a valuable testbed for controlled analysis because its sequences are labeled from a neutral onset to a peak expression frame [2]. Survey literature consistently highlights cross dataset generalization, class imbalance, subtle or ambiguous expressions, and unconstrained acquisition conditions as open challenges for the field [10].

III. MATERIALS AND METHODS

A. Dataset

The experiments in this study are conducted on CK+, an extended version of the original Cohn Kanade dataset. CK+ comprises 593 image sequences collected from 123 participants, of which 327 sequences carry validated emotion labels together with Facial Action Coding System (FACS) annotations. Seven emotion categories are considered: anger, disgust, fear, happiness, sadness, surprise, and neutral. Because CK+ is recorded under controlled illumination with predominantly frontal head poses, it offers a reproducible basis for model comparison, although it does not fully capture the variability encountered in unconstrained, real world deployment [2]. Table I summarizes the principal characteristics of the dataset.

TABLE I - CK+ Dataset Characteristics

Attribute	Value
Participants	123

Natural Resources for Human Health
eISSN: 2583-1194

Attribute	Value
Total sequences	593
Emotion labeled sequences	327
Emotion classes	7
Sequence structure	Neutral to peak expression
Acquisition conditions	Controlled illumination; mainly frontal pose

B. Preprocessing and Data Augmentation

Every image is passed through a common preprocessing chain consisting of face detection, cropping and alignment of the facial region, resizing to the input dimensions expected by each network, pixel value normalization, and light quality enhancement. During training, the pipeline additionally applies horizontal flipping, small angle rotation, affine warping, and color jitter. These augmentations broaden the range of head pose, position, and illumination that the network observes during training without altering the underlying emotion label [11]. Applying the same augmentation protocol to all three architectures removes preprocessing as a confounding variable and ensures that the reported performance differences reflect model design rather than data handling.

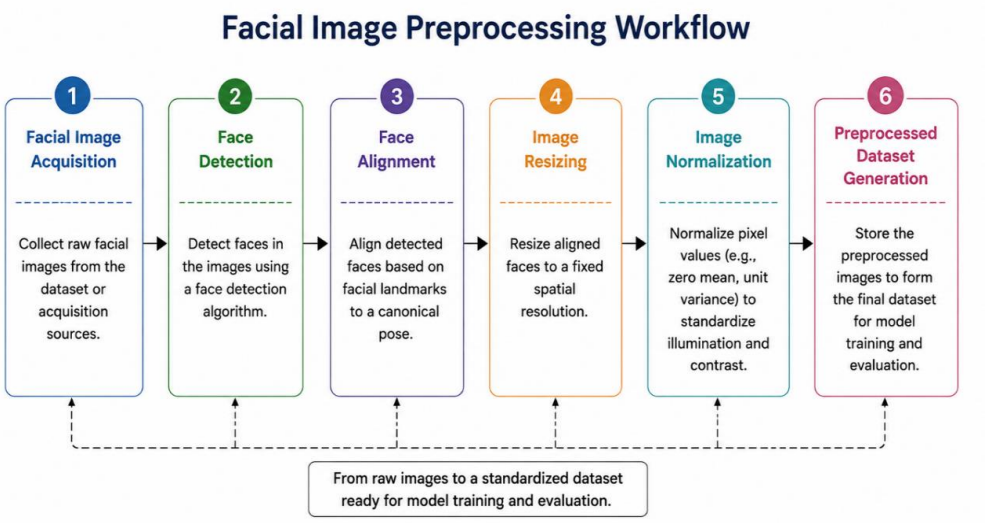

Fig.1. Facial image preprocessing workflow applied prior to model training and evaluation.

C. Evaluation Metrics

Recognition performance is quantified using accuracy, precision, recall, and F1 score. For a given class with true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), these quantities are computed as accuracy = (TP + TN) / (TP + TN + FP + FN), precision = TP / (TP + FP), recall = TP / (TP + FN), and F1 = 2 × precision × recall / (precision + recall). In addition to classification quality, the computational profile of each model is characterized in terms of trainable parameter count, multiply–accumulate operations (FLOPs), estimated memory footprint, and qualitative suitability for real time or edge deployment.

Natural Resources for Human Health
eISSN: 2583-1194

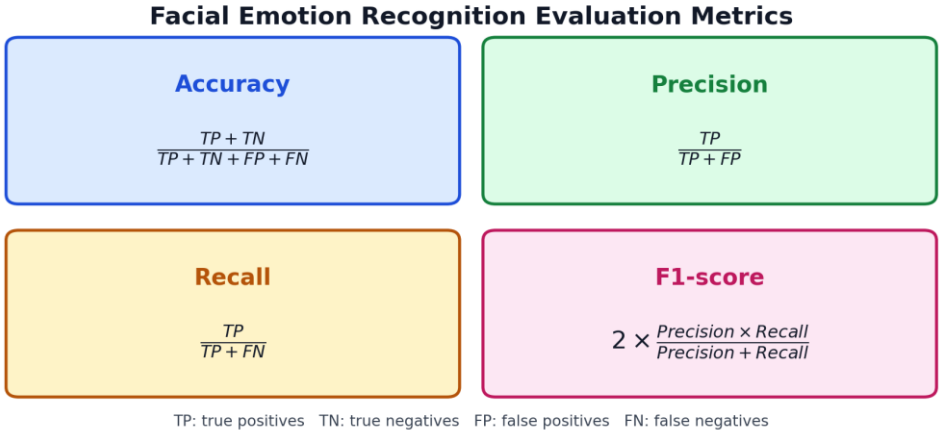

Fig.2. Evaluation metrics used to assess facial emotion recognition performance.

IV. MODEL ARCHITECTURES

A. Custom CNN Baseline

The custom CNN is a compact baseline architecture built from repeated convolution, ReLU activation, batch normalization, and max pooling stages, followed by dense layers and a seven way softmax output. The earlier convolutional stages capture low level cues such as edges and local texture, while the deeper stages learn configurations more directly tied to facial expression. The principal advantage of this design is its very low computational cost and memory footprint, which make it attractive for embedded inference; its limited representational capacity, however, can reduce recognition performance for subtle classes such as fear and sadness.

B. VGG16 Transfer Learning

The VGG16 branch is initialized with pretrained visual features and adapted for FER by replacing its original classification head with task specific dense and softmax layers. Fine tuning allows generic low and mid-level features edges, contours, textures, and shapes to specialize toward facial expression cues. This adaptation yields stronger recognition accuracy than the custom CNN, but at a considerably higher cost: the model carries 134.29 million parameters and requires 11.69 GFLOPs, making it the most computationally demanding architecture examined in this study.

C. Proposed EmotionNet

EmotionNet is composed of stacked convolutional blocks in which channel depth progressively increases from 64 to 512 while the spatial resolution of the feature maps is correspondingly reduced. Each block integrates convolution, batch normalization, ReLU activation, max pooling, and dropout, allowing the network to learn increasingly abstract expression related representations while limiting internal covariate shift and overfitting. The resulting high level features are flattened and passed through fully connected layers before a seven class softmax layer produces the final prediction.

Training minimizes categorical cross entropy using the Adam optimizer. Each iteration performs mini batch forward propagation, loss computation, backpropagation, and parameter updates, with validation performance monitored after every epoch. Learning rate scheduling, early stopping, dropout, and batch normalization are used jointly to stabilize training. Under the selected configuration, EmotionNet converges in approximately 45 epochs, with an average epoch duration of 24.8 s.

D. Real Time Video Inference Workflow

EmotionNet is embedded in a deployment pipeline that accepts either a live camera feed or a prerecorded video file. The pipeline performs video acquisition, frame extraction, face detection, face preprocessing, EmotionNet inference, frame annotation, and reconstruction of the output stream. For each detected face, the predicted emotion label and its associated confidence score are overlaid using a bounding box, and annotated frames are reassembled in their original temporal order. This design supports continuous, low latency visual feedback suitable for monitoring and interactive applications. The stages of this workflow are summarized in Table II and illustrated in Fig. 3.

Natural Resources for Human Health
eISSN: 2583-1194

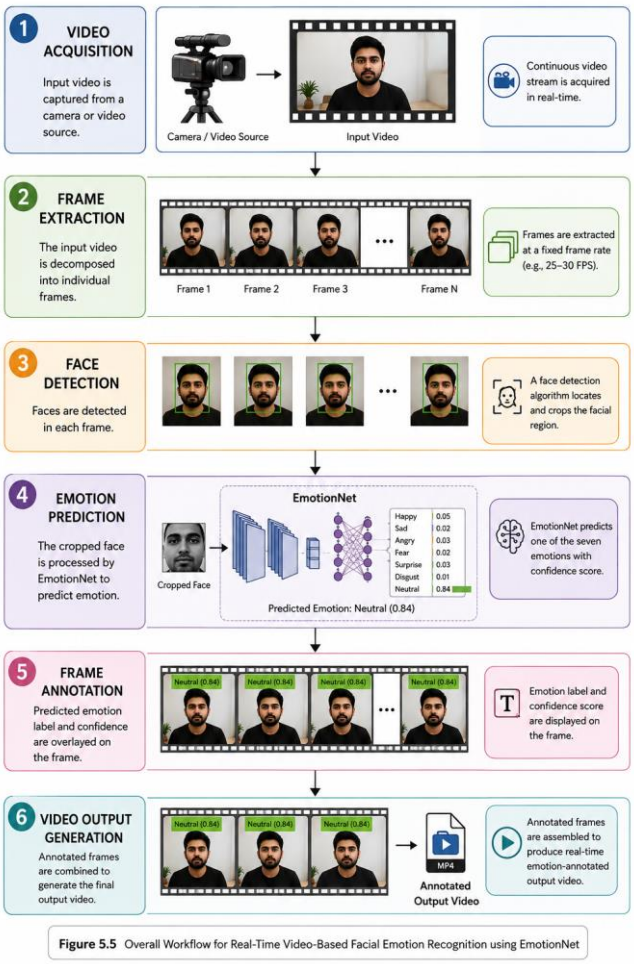

Fig.3.End to end real time video based facial emotion recognition workflow using EmotionNet.

TABLE II - Real Time Video Processing Workflow

Stage	**Primary Function**
Acquisition	Capture live or stored video
Frame extraction	Convert video stream into individual frames
Face detection	Localize and crop the facial region
Emotion prediction	Classify one of seven emotion categories
Annotation	Overlay predicted label and confidence score
Output generation	Reconstruct the annotated video stream

V. EXPERIMENTAL RESULTS

A. Overall Recognition Performance

Table III summarizes the classification performance of the three architectures. EmotionNet attains the highest value on every metric, exceeding the custom CNN by 6.02 percentage points and VGG16 by 2.13 percentage points in accuracy; the corresponding F1 score improvements are 6.47 and 2.36 percentage points, respectively. This pattern indicates that the deeper, task specific representation learned by EmotionNet improves both overall class separability and the balance between precision and recall.

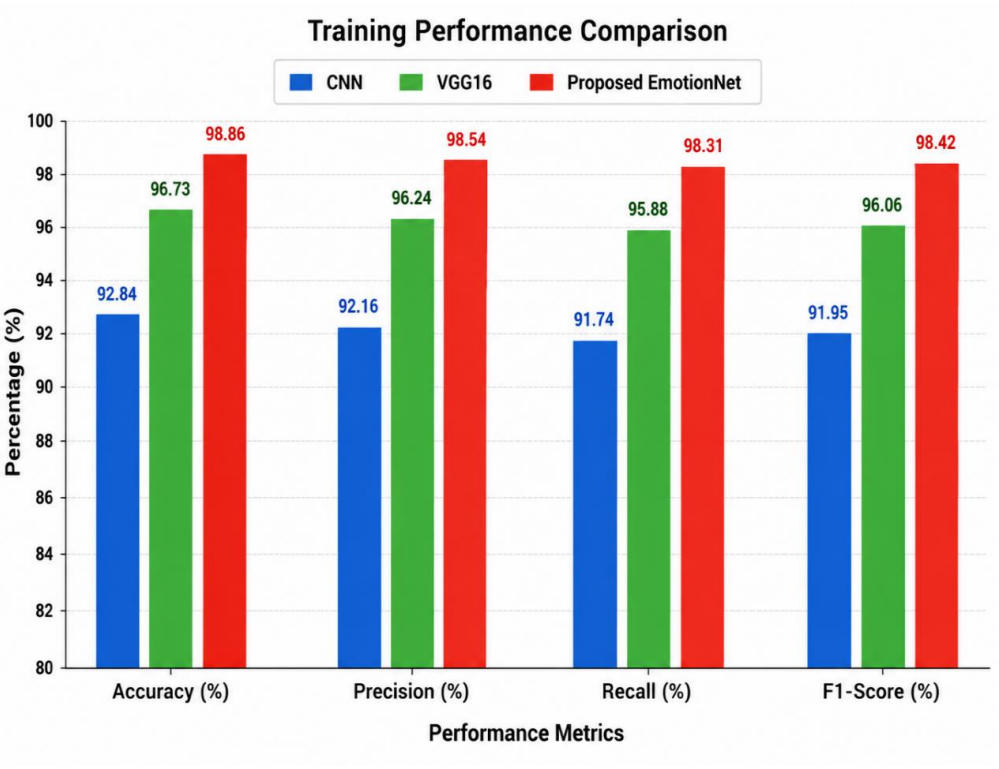

Fig.4. Accuracy, precision, recall, and F1 score comparison across the three models.

TABLE III - Overall Classification Performance (%)

Model	Accuracy	Precision	Recall	F1 score
Custom CNN	92.84	92.16	91.74	91.95
VGG16	96.73	96.24	95.88	96.06
EmotionNet	98.86	98.54	98.31	98.42

B. Emotion Wise Performance of EmotionNet

At the per class level, happiness is recognized most reliably, with 99.84% accuracy and a 99.81% F1 score, followed closely by surprise and disgust. Fear is the most challenging category, achieving 97.24% accuracy and a 97.38% F1 score, because its characteristic widened eyes and open mouth partially overlap with the visual signature of surprise. Sadness is likewise comparatively difficult to separate, as its cues are often subtle and can resemble a neutral expression. Nevertheless, every

emotion category exceeds 97% accuracy, indicating that EmotionNet maintains consistently strong recognition across all seven classes. Detailed per class results are reported in Table IV and Fig. 5.

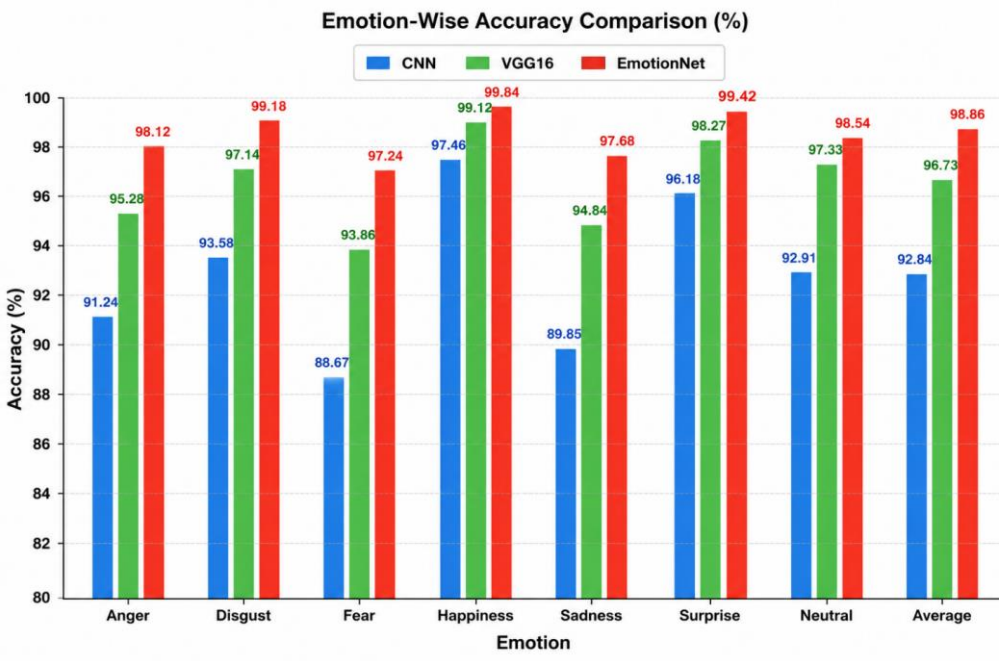

Fig.5. Emotion wise accuracy comparison for the custom CNN, VGG16 and EmotionNet.

TABLE IV - Emotion Wise Performance of Emotionnet (%)

Emotion	Accuracy	Precision	Recall	F1 score
Anger	98.12	98.34	97.88	98.11
Disgust	99.18	99.26	99.05	99.15
Fear	97.24	97.82	96.94	97.38
Happiness	99.84	99.91	99.72	99.81
Sadness	97.68	98.04	97.46	97.75
Surprise	99.42	99.46	99.38	99.42
Neutral	98.54	98.95	98.12	98.53
Average	98.86	98.54	98.31	98.42

C. Convergence and Training Behavior

EmotionNet exhibits steady and stable convergence over training. Training and validation accuracy rise from 82.45% and 80.92%, respectively, during epochs 1–10, to 99.42% and 98.86% during epochs 41–50, while validation loss decreases from 0.624 to 0.042 over the same interval. The small final gap between training and validation accuracy, together with stable gradient

behavior and low validation loss, suggests that the combination of dropout, batch normalization, and early stopping effectively controls overfitting. The epoch wise progression is reported in Table V and Fig. 6.

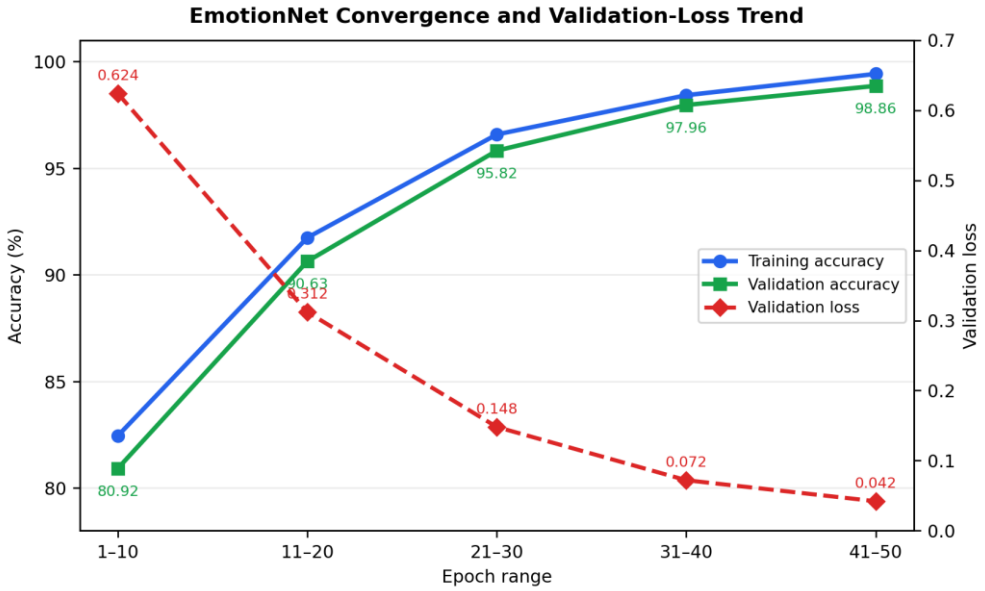

Fig. 6. EmotionNet training and validation accuracy convergence with validation loss reduction.

TABLE V - Emotionnet Convergence Behavior

Epoch Range	Train Acc. (%)	Val. Acc. (%)	Val. Loss
1–10	82.45	80.92	0.624
11–20	91.74	90.63	0.312
21–30	96.58	95.82	0.148
31–40	98.42	97.96	0.072
41–50	99.42	98.86	0.042

D. Computational Complexity and Memory Trade offs

The custom CNN is the least computationally demanding model, requiring only 62.50 MFLOPs and an estimated 3.64 MB of memory. EmotionNet requires more resources than the custom CNN but remains substantially lighter than VGG16 across every computational dimension. Relative to VGG16, EmotionNet reduces the parameter count by approximately 68.2%, computation by 33.1%, and estimated total memory by 48.7%, while simultaneously improving accuracy by 2.13 percentage points. This combination positions EmotionNet as the most balanced architecture for high accuracy, real time use cases, whereas the custom CNN remains the preferred option when memory and compute budgets are the dominant design constraints. Table VI and Fig.7 present the full architectural and computational comparison.

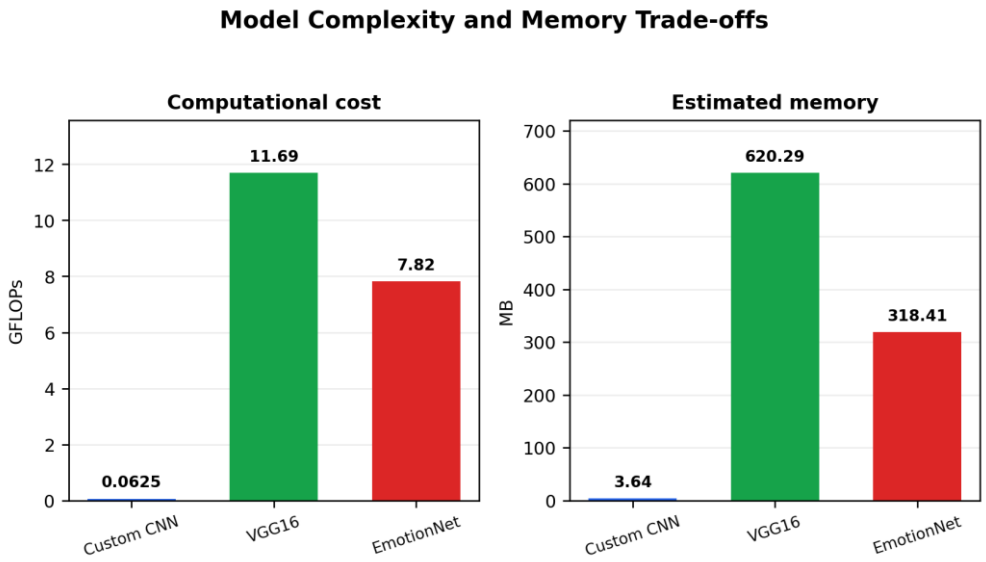

Fig. 7. Computational cost and estimated memory comparison across the three models.

TABLE VI Architectural And Computational Comparison

Metric	Custom CNN	VGG16	EmotionNet
Total parameters	94,551	134,289,223	42,704,711
Trainable parameters	94,551	119,574,535	42,704,711
Non trainable parameters	0	14,714,688	0
Computation	62.50 MFLOPs	11.69 GFLOPs	7.82 GFLOPs
Estimated memory	3.64 MB	620.29 MB	318.41 MB

VI. DISCUSSION

The results outline a clear accuracy efficiency spectrum across the three architectures. The custom CNN offers fast, lightweight inference but has limited capacity to learn sufficiently rich facial representations. VGG16 improves recognition accuracy through transfer learning, yet its parameter count and memory overhead constrain its use on edge hardware. EmotionNet occupies an intermediate computational tier while achieving the highest recognition scores among the three models; its gains can be attributed to progressive channel expansion, repeated normalization and nonlinear transformation, controlled spatial down sampling, and dropout based regularization.

Across the baseline models, misclassifications occur predominantly between fear and surprise, sadness and neutral, and anger and disgust—pairs of emotions that share overlapping facial cues or differ only through subtle muscular movement. EmotionNet reduces, though does not fully eliminate, this confusion. Incorporating temporal aggregation across video frames, facial action unit supervision, and attention mechanisms focused on discriminative eye and mouth regions may further improve recognition of these difficult classes.

Three limitations should be considered when interpreting these results. First, CK+ is captured under controlled conditions and consists mainly of posed expressions, so high within dataset accuracy may not transfer directly to spontaneous, real world

behavior. Second, the demographic diversity of the dataset is moderate. Third, the computational assessment reported here is based primarily on model level parameter counts and FLOPs; a complete characterization of latency and energy consumption across diverse edge hardware would be required before operational deployment.

VII. APPLICATIONS AND DEPLOYMENT CONSIDERATIONS

EmotionNet can support a range of applications, including healthcare observation, adaptive learning environments, driver state monitoring, intelligent surveillance, social robotics, entertainment, and human computer interaction more broadly. Responsible deployment in these settings should incorporate informed consent, privacy safeguards, auditing of performance across demographic subgroups, confidence thresholds for automated decisions, and human review in consequential applications. It is important to note that FER estimates an individual's visible facial expression rather than their complete internal emotional state; consequently, model predictions should be treated as contextual signals that inform, rather than replace, human judgment, particularly in sensitive or high stakes settings.

VIII. CONCLUSION AND FUTURE WORK

This paper presented a controlled comparison of a custom CNN, a VGG16 transfer learning model, and the proposed EmotionNet architecture for seven class facial emotion recognition on the CK+ dataset. EmotionNet achieved the strongest performance overall, with 98.86% accuracy, 98.54% precision, 98.31% recall, and a 98.42% F1 score, outperforming both baseline models while also reducing computational cost and estimated memory usage relative to VGG16. These results indicate that EmotionNet offers a favorable balance between recognition quality and real time feasibility. The custom CNN remains the more appropriate choice for severely resource constrained edge devices, while EmotionNet is preferable whenever recognition accuracy and scalable deployment are the primary design objectives.

Future work should evaluate EmotionNet on spontaneous, in the wild datasets and conduct cross dataset generalization testing, integrate temporal modeling through recurrent or transformer based modules, and investigate model compression techniques such as quantization, pruning, and knowledge distillation for deployment on constrained hardware. Incorporating explainable attention or saliency based mechanisms, along with fairness aware evaluation across demographic groups, would further improve the transparency and responsible adoption of the proposed system.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," J. Pers. Soc. Psychol., vol. 17, no. 2, pp. 124–129, 1971.

[2] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn Kanade dataset (CK+): A complete dataset for action unit and emotion specified expression," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2010, pp. 94–101.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Adv. Neural Inf. Process. Syst., vol. 25, 2012, pp. 1097–1105.

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 448–456.

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510–4520.

[10] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Affect. Comput., vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022.

[11] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," J. Big Data, vol. 6, art. no. 60, 2019.

[12] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.

[13] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd sourced label distribution," in Proc. ACM Int. Conf. Multimodal Interaction, 2016, pp. 279–283.

[14] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Netw., vol. 64, pp. 59–63, 2015.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.