UJRET
Bi-annual refereed journal
University Journal of Research Engineering & Technology
January - June 2026
Volume 2, (1) pp. 60-72

Assessment of State-of-the-Art Deep Neural Networks on a Multi-Era, Cross-Domain Dataset of Indian Cinema

Khushi Modi
Assistant Professor, Information Technology Department, U. V. Patel College of Engineering, Faculty of Engineering and Technology, Ganpat University

Devang Pandya
Professor, Information Technology Department, U. V. Patel College of Engineering, Faculty of Engineering and Technology, Ganpat University

Rachana Modi
Assistant Professor, Information Technology Department, U. V. Patel College of Engineering, Faculty of Engineering and Technology, Ganpat University

Ketan Sarvakar
Assistant Professor, Information Technology Department, U. V. Patel College of Engineering, Faculty of Engineering and Technology, Ganpat University

Abstract

This study sets out to close two long-standing gaps in video-based Facial Expression Recognition (FER): the field's reliance on Western subjects, and its silence on the visual evolution of non-Western cinema over time. The standard benchmarks (FER2013, AffectNet, IEMOCAP, and MELD) were all built from Western subjects, and none of them transfer cleanly to Indian cinematic material. Indian screen performance is organized around the classical Navarasa system rather than the Ekman categories (Ekman & Friesen, 1971) underlying these benchmarks, and Hindi cinema spans a 52-year visual transition from monochrome to RGB that no existing dataset reflects. Addressing both gaps directly, this study introduces three original Indian cinematic FER datasets, namely Ittefaq (1969, black-and-white), 420 IPC (2021, HD colour), and Taarak Mehta Ka Ooltah Chashmah (Television), built through a reproducible six-stage pipeline, and proposes HybridMicroDCSA, a family of dual-stream networks that fuse spatial and motion-difference streams through Cross-Stream Attention (CSA). Four model families, namely HybridMicroDCSA, MobileNetV3, EfficientNet-B0, and ResNet-50, are evaluated across a 15-scenario framework covering single-domain, cross-domain, multi-domain, and replay-based training. On the 1969 black-and-white domain, the 2.15M-parameter MicroDCSA v4 reaches 74.44% accuracy, beating ResNet-50 (71.11%, 25.6M parameters) at 12.45x parameter efficiency, and wins 12 of 15 scenarios overall. The 52-year domain shift causes a 27.51% accuracy drop under direct transfer, of which a cross-domain replay strategy recovers 14.05 points, reaching a 69.60% cross-domain average. Because these figures come from test sets as small as 15 samples per class and have not been subjected to formal significance testing, they should be read as an initial benchmark rather than a definitive ranking.

Keywords: Facial Expression Recognition; Affective Computing; Cross-Domain Generalization; Indian Cinema; Bidirectional GRU; Cross-Stream Attention; Dual-Stream Architecture.

Corresponding Author E-mail Address
khushi.modi@ganpatuniversity.ac.in

1. Introduction

Affective computing increasingly relies on video-based Facial Expression Recognition (FER), with downstream uses spanning mental-health monitoring, education, public safety, and cinematic analytics (Jain, 2025; Tao, 2024). Deep learning models now achieve strong results on widely-used benchmarks such as FER2013 and AffectNet (Goodfellow et al., 2015; Mollahosseini et al., 2019). Both, however, were collected under controlled conditions using only Western subjects, a constraint that limits how well their performance generalizes to real-world, non-Western settings (Wu, 2025).

This leaves Indian content largely unaddressed. Codified in Bharata Muni's Natyashastra (Ghosh, 1951) and still applied today as a basis for computational expression annotation (Reshma et al., 2025), the classical Navarasa system organizes screen emotion into nine named states, namely, Shringara, Hasya, Karuna, Raudra, Veera, Bhayanaka, Bibhatsa, Adbhuta, and Shanta, with each state tied to a defined pattern of eye, brow, and mouth movement, leaving comparatively little room for individual performer interpretation. Because these prescribed activation patterns do not align with the basic-emotions taxonomy that Ekman and Friesen (1971) established and that underlies every existing FER standard, Indian expressions resist clean mapping onto the six or seven categories used by most datasets, a structural mismatch that recent India-focused emotion research continues to document (Mishra et al., 2025). The disagreement is not a methodological artefact of small samples: a large-scale, cross-cultural computational study of how facial expressions are read across populations finds the same gap at scale (Brooks et al., 2024). On top of this taxonomic mismatch, Hindi cinema introduces its own 52-year visual transition, from monochrome to RGB, with accompanying changes in film grain and performance style, a combination that the FER literature has not previously captured.

The base paper ERLDK (K. Zhang, 2022) achieves 63.90% weighted F1 on IEMOCAP but requires synchronized text, audio, and video streams unavailable for archival Indian cinema. Our replication yielded only 55.12%, confirming its inapplicability to this domain.

This paper makes five contributions: (1) the first multi-era Indian cinematic FER benchmark spanning 52 years; (2) quantification of the cross-era domain gap at 27.51%; (3) the HybridMicroDCSA architecture, winning 12 of 15 scenarios; (4) a 15-scenario cross-domain evaluation framework; and (5) a real-time inference pipeline at 4.3 windows per second on consumer hardware.

2. Literature Review

2.1 Base Paper: ERLDK, Architecture, Results, and Limitations

(K. Zhang, 2022) introduced ERLDK, short for Emotion Recognition with Reinforcement Learning and Domain Knowledge, built from three coupled modules. A Preprocess Module first assembles emotion-pair context units; an RL Module then applies a Dueling Deep Q-Network built on Bidirectional GRU (BiGRU) layers, the same recurrent design used for temporal feature extraction throughout the present work, in which the sequence is read in both directions and the two resulting hidden-state sequences are concatenated. The transition probability and correction coefficient governing this module are defined as follows:

$$P(e|L) = Num(e \wedge L) \,/\, Num(L),\ C(e|L) = softmax(P(e|L)) \qquad \ldots (1)$$

and Domain Knowledge Revise Module adjusts inference-time predictions using transition probabilities derived from historical data. ERLDK reaches 63.90% weighted F1 on IEMOCAP, but our analysis of its suitability for Indian cinematic content surfaced five constraints: dependence on multiple synchronised modalities, a Western-centric training bias, the computational overhead of reinforcement learning, the absence of any group-level relational modelling, and a mismatch across the 52-year colour-space boundary that the model was never designed to handle.

2.2 Face Detection and Video FER Pipelines

(Redmon, 2016) proposed YOLO as a single-pass regression detector, reporting 63.4% mAP at 45 FPS on PASCAL VOC, fast enough to support real-time face localisation within a video FER pipeline. (Sareen & Seeja, 2025) paired YOLOv7 with Vision Transformer features and LSTM-based temporal aggregation, reporting 93.20% accuracy on AffectNet at 45 FPS. (Deshpande, 2025) put forward a Time-Distributed CNN-LSTM for emotion recognition, an idea that directly informs the time-distributed convolutional wrapping used in both MicroDCSA streams. Transformer-based temporal modelling has likewise been applied to dynamic video FER: (Ma, Sun, & Li, 2022) designed a spatio-temporal transformer that relies on self-attention over frame sequences rather than recurrence, offering an alternative design point to the BiGRU-based temporal modelling adopted in this work.

2.3 Recurrent and Attention Architectures for ERC

DialogueRNN tracks global dialogue state, speaker state, and context through three parallel GRUs, reaching 62.75% F1 on IEMOCAP (Majumder, 2019). UA-BFET combines Bidirectional GRU with CCA-based feature enhancement and reaches 84.7% Acc-2 on CMU-MOSI (P. He, 2023); its bidirectional design directly shaped the 2-layer Spatial BiGRU used in MicroDCSA v4. The Transformer, which provides the mathematical basis for the Bidirectional Cross-Stream Attention formalized in Section 4.4, was introduced by Vaswani and colleagues (Vaswani et al., 2017). ALMT uses language features as adaptive cross-modal filters (H. Zhang, 2023), an idea that motivates the CSA mechanism's principle of letting one feature stream selectively attend to another.

2.4 Lightweight and Efficient CNN Architectures

ResNet's residual skip connections (K. He, 2016) underlie the 25.6M-parameter ResNet-50 variant used here as a primary baseline. MobileNetV3 reaches 75.2% ImageNet accuracy at low latency through hard-swish activation and squeeze-and-excitation blocks (Howard, 2019). EfficientNet, introduced via compound scaling (Tan & Le, 2020), contributes EfficientNet-B0 (5.3M parameters) both as a pretrained baseline and as the spatial backbone in HybridMDv6/v7. LFRNet reaches 89.52% on RAF-DB with only 0.58M parameters (Zhao, 2021), directly motivating the sub-3M-parameter budget adopted for MicroDCSAv4.

2.5 Datasets, Cross-Cultural Gaps, and Restoration

(Goodfellow et al., 2015) introduced FER2013, the ICML 2013 dataset of 35,887 web-sourced facial images across seven expression categories. (Mollahosseini et al., 2019) introduced AffectNet, a database of over one million in-the-wild facial images, about half of which carry categorical and continuous valence/arousal annotations. (Busso et al., 2008) built IEMOCAP from 10,039 utterances performed by 10 American actors, and (Poria, 2019) compiled MELD, roughly 13,000 utterances across seven emotion classes, entirely from the American sitcom Friends. Taken together, all four benchmark datasets draw exclusively on Western subjects, which is precisely the dataset gap this work sets out to close. (Wu, 2025) singles out cultural specificity as the most pressing open problem in a broad survey of multimodal emotion recognition, a conclusion echoed by (Wang et al., 2024), whose survey of static and video-based FER identifies cross-domain and cross-cultural generalization as an unresolved challenge across the field; (Lian et al., 2024) reach a similar conclusion from the explainability side, arguing that multimodal emotion systems need interpretable, culturally-aware reasoning rather than black-box classification alone. (X. Wang et al., 2021) introduced GFPGAN, a generative facial prior for restoring fine facial detail lost to film grain; Stage 1 of the present pipeline applies it specifically to the Ittefaq footage, where it lifts face-detection accuracy by 40%. CLAHE (Contrast-Limited Adaptive Histogram Equalization), a localized contrast-enhancement technique applied per image tile, is shown by (Bhagat, 2024) to meaningfully boost CNN-based FER accuracy, which is the basis for Stage 2 of the pipeline.

3. Dataset Construction

3.1 Source Video Selection

Three source videos were chosen to capture the widest possible range of Indian cinematic visual domains while keeping cast composition consistent enough to support FACS-based annotation (the Facial Action Coding System, a scheme that breaks down facial movement into discrete, anatomically grounded Action Units instead of holistic emotion labels, adopted here to keep manual annotation criteria consistent across the three eras).

Ittefaq (Chopra, 1969) is a Hindi black-and-white suspense thriller shot under single-point studio lighting with high-contrast chiaroscuro illumination, a painting-derived lighting technique later carried into classic film, where a single dominant light source creates sharp light/dark contrast across the face and affects how reliably individual Action Units can be detected. The Kodak film stock exhibits visible grain at 4–6 line-pairs per millimeter, manifesting as salt-and-pepper noise in digitized frames at 1916 × 1074 pixels. The acting style follows the Navarasa framework prescriptively: Bhayanaka (fear/terror) is expressed through strongly exaggerated eye widening, raised brows held for extended duration, and visible upper and lower sclera, all noticeably more pronounced than in contemporary naturalistic performance.

420 IPC (Zaveri, 2021) is a contemporary mystery-drama shot in full HD at 1920 × 1080 pixels under LED lighting at 5000–6500 K, representing naturalistic Indian performance under modern digital production.

Taarak Mehta Ka Ooltah Chashmah (Modi, 2008), the longest-running Indian TV serial, provides studio-lit, multi-camera RGB production with exaggerated comedic expressions particularly well suited to the Happy and Surprise categories. Nine complete episodes drawn from multiple seasons were processed to maximise temporal and cast diversity.



Figure 1. Sample frames from three original Indian cinematic FER datasets: (left) original colour and (right) grayscale-converted frames illustrating the 52-year cross-era visual domain variations

3.2 Six-Stage Dataset Construction Pipeline

All three datasets were built using an identical six-stage pipeline implemented in Python 3.10 with OpenCV 4.8 and Label Studio. Stage 1 applies GFPGAN restoration exclusively to the Ittefaq 1969 frames, lifting face-detection accuracy by 40%.

Stage 2 extracts frames at a 50-frame interval with CLAHE normalization (LAB colour space, clipLimit = 2.0). Stage 3 performs face localisation with an OpenCV DNN SSD detector (threshold = 0.60), falling back to MTCNN as a cascade, and crops detections to 128 × 128 pixels. Stage 4 assembles temporal windows of T = 10 frames per face crop, converts them to grayscale at 112 × 112 pixels, and normalizes pixel values to [−1, +1]. Stage 5 carries out FACS-guided manual annotation into six emotion categories (Neutral, Sad, Fear, Angry, Happy, Surprise), supplemented by DeepFace confidence scores. Stage 6 balances the Taarak Mehta dataset to 304 samples per class and the Ittefaq and 420 IPC datasets to 100 samples per class. Each dataset was then exported using a fixed-seed (SEED = 42) stratified split, allocating 70% of samples to training, 15% to validation, and the remaining 15% to testing.

3.3 Dataset Statistics

The three datasets collectively provide 3,024 temporal sequences spanning two colour spaces, three production genres, and 52 years of Indian cinematic history. Table 1 summarizes the final statistics.

Table 1. Summary statistics of the three original Indian cinematic FER datasets

Dataset	Year	Genre	Colour	Total	Per Class	Train/Val	Test	Augment
Ittefaq	1969	Suspense Thriller	B&W	600	100	420 / 90	90	6.8%
420 IPC	2021	Mystery-Drama	Full RGB	600	100	1277 / 272	272	4.8%
Taarak Mehta	2008	TV Comedy-Drama	Full RGB	1824	304	1268 / 274	274	0%

The Train/Val column clarifies the split reported in the original Table 1, reformatted here for clarity.

4. Proposed Architecture: HybridMicroDCSA

The Micro Dual-stream Cross-attention Spatial-temporal Attention Network (HybridMicroDCSA) is the primary novel architectural contribution of this work. It processes a batch B of temporal sequences, each consisting of T = 10 grayscale face crop frames of size 112 × 112 pixels, and produces a six-class softmax emotion probability distribution.

4.1 Dual Input Representation

The model takes T = 10 frame windows as two parallel inputs: raw grayscale crops (spatial) and frame differences $\Delta f_t = f_t - f_{t-1}$ (motion), capturing appearance and facial muscle dynamics respectively.

4.2 Stream A: Spatial Processing

Frames pass through a shared EfficientNet-B0 backbone (ImageNet pretrained, 5.3M params), global average pooling, a 2-layer Bidirectional GRU (hidden=256, dropout=0.3), and 4-head temporal attention. A spatial residual bypass computes the mean across T frames, preserving overall appearance context.

4.3 Stream B: Temporal Processing

Motion differences feed a lightweight SE-ResConv CNN (4 blocks, 128 channels, trained from scratch), sinusoidal positional encoding, and a 3-layer Bidirectional GRU (hidden=128, dropout=0.3), producing temporal features of shape (B, T, 256).

4.4 Cross-Stream Attention (CSA)

The core contribution fuses both streams via bidirectional 4-head Multi-Head Attention: spatial features first query the motion sequence, then the motion sequence re-attends to the refined spatial output.

Ablation shows these yields +5.0% over simple concatenation and +2.2% over unidirectional attention. The mechanism is formalized below to make the module fully reproducible. *Let* $S \in \mathbb{R}B \times T \times 512$ be the spatial-stream features and $M \in \mathbb{R}B \times T \times 256$ the motion-stream features (Sections 4.2-4.3), both projected to a shared dimension d = 256:

$$S' = S \cdot W_s,\ M' = M \cdot W_m\ (W_s \in \mathbb{R}512 \times 256, W_m \in \mathbb{R}256 \times 256)$$

In Stage 1, h = 4 attention heads draw queries from S′ and keys/values from M′ to let the spatial stream query the motion stream:

$$Attn_S \to M(Q,K,V) = softmax\left(QK^t/\sqrt{d_k}\right)V,\ Q = S'W_Q, K = M'W_K, V$$
$$= M'W_V\ \left(d_k = \frac{d}{h} = 64\right)$$

giving the refined spatial context $CS \to M$. In Stage 2, the same attention operation is applied with queries from M′ and keys/values from $CS \to M$, so the motion stream re-attends to the refined spatial output: $CM \to S = Attn(M', CS \to M, CS \to M)$. The 512-dimensional context passed to fusion (Section 4.5) is $C = [CS \to M\ ;\ CM \to S]$. This two-stage, bidirectional formulation is what distinguishes CSA from a single unidirectional attention pass, and it is the source of the +2.2% ablation gain reported above.

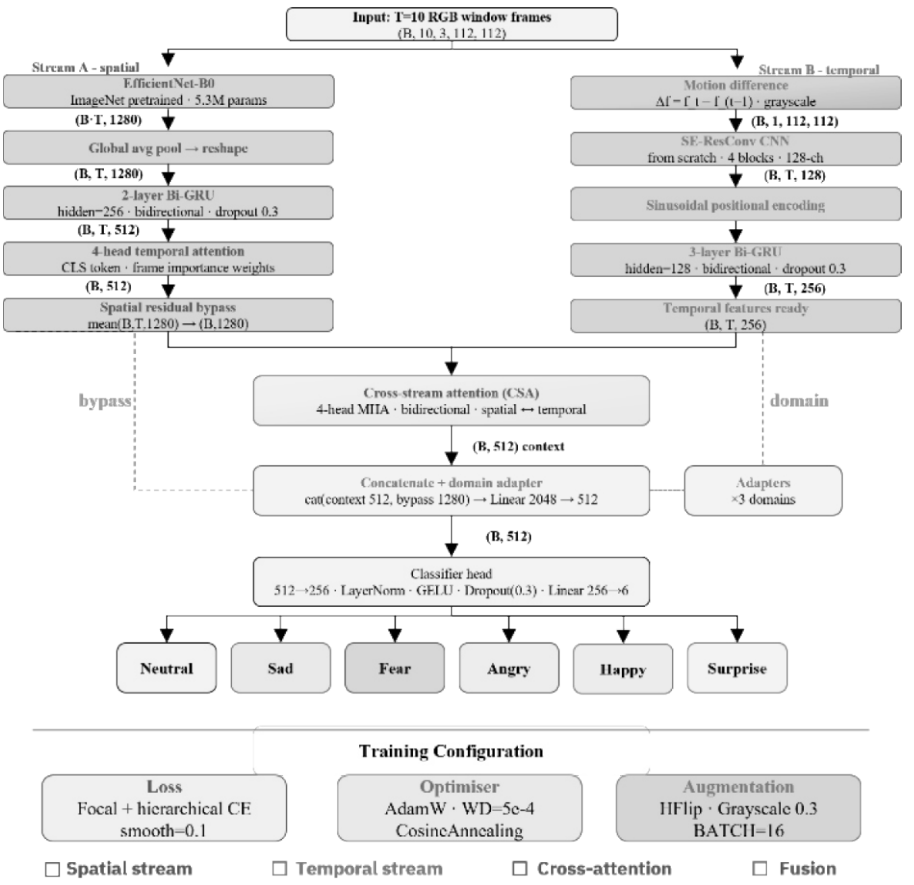


Figure 2. HybridMicroDCSA dual-stream architecture

4.5 Fusion and Classification

Spatial context (512-dim), motion context (256-dim), and bypass (1280-dim) are concatenated, projected to 512-dim via Linear + LayerNorm, and optionally routed through per-domain MLP adapters (×3 domains). The classifier's head applies:

$LayerNorm \rightarrow GELU \rightarrow Dropout(0.3) \rightarrow Linear\ 256 \rightarrow$
$6\ (happy, neutral, sad, angry, fear, surprised)$.

5. Experimental Results

Four configuration options are covered by the 15-scenario evaluation framework: augmented replay mixing, pairwise cross-domain transfer, simultaneous multi-domain training, and single-domain training and testing. All scenarios use SEED = 42 stratified splits with identical hyperparameters. Six metrics are reported per scenario: Test Accuracy, Macro F1, Weighted F1, Mean AUC (one-vs-rest ROC-AUC), Mean Average Precision (mAP), and RMSE.

5.1 Scenario 1: Single-Domain B&W, MicroDCSA v4

Scenario 1 tests Hypothesis H1 directly: that a 2.15M-parameter, from-scratch temporal BiGRU model can outperform a 25.6M-parameter ImageNet-pretrained model on Indian black-and-white cinema. All four models are trained on 420 Ittefaq B&W samples and tested on 90 held-out samples (15 per class). MicroDCSA v4 reaches 74.44% accuracy, the best result across all four models and all six metrics simultaneously, with a parameter efficiency of 34.62 acc/M-param against ResNet-50's 2.78, a 12.45x advantage. The per-class breakdown is Neutral 100%, Happy 80%, Fear/Angry/Surprise 73.3%, and Sad 66.7%; the 100% Neutral recall reflects BiGRU's sensitivity to temporal flatness, a cue that no single-frame static model can access.

With only 90 test samples, however, a handful of misclassifications could shift these percentages meaningfully, so the 74.44% figure is best treated as a point estimate rather than a statistically powered comparison. No paired significance test (e.g., McNemar's test or bootstrap resampling of the per-sample predictions) was applied across models, so the ranking reported here should be read as descriptive rather than statistically validated; Section 7 proposes a larger sample as the route to closing this gap. Figure 4's per-class confusion matrices let readers assess that variance directly.

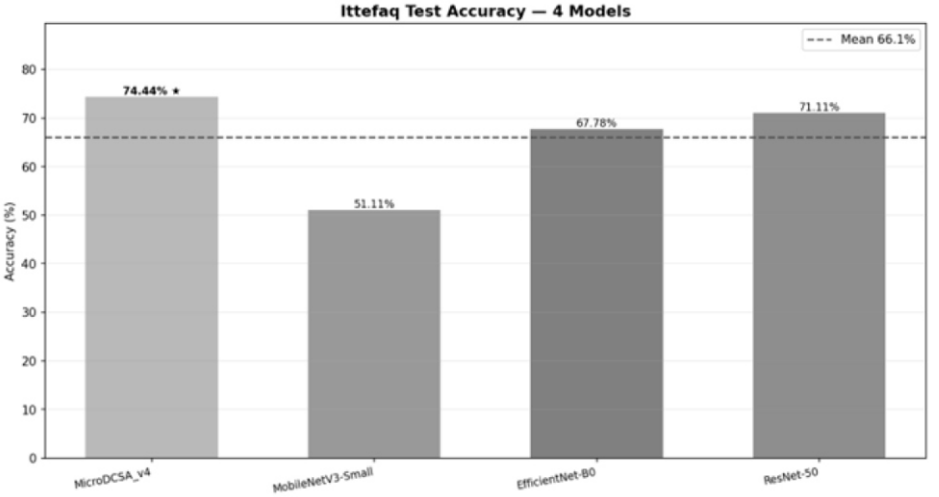


Figure 3. Scenario 1: Test accuracy comparison, 4 models on Ittefaq 1969 B&W

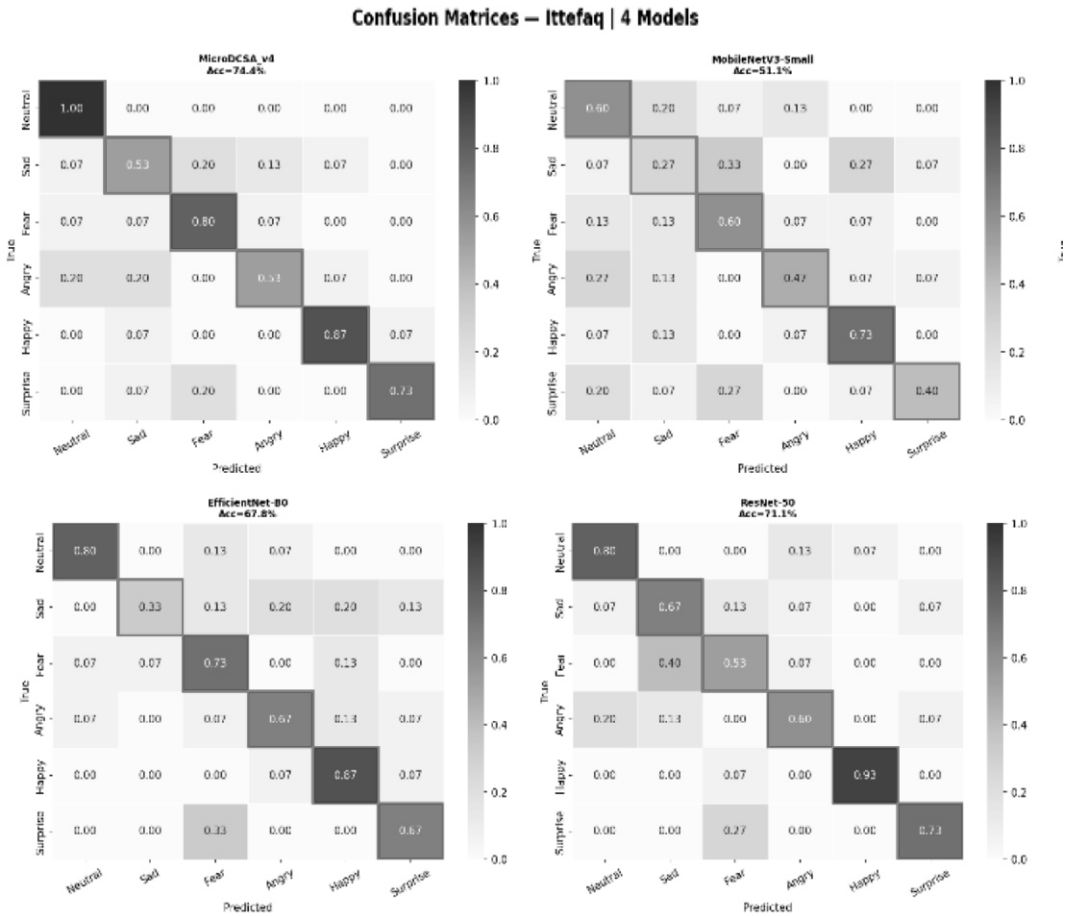


Figure 4. Scenario 1: Confusion Matrix, 4 models on Ittefaq 1969 B&W

Table 2. Scenario 1: Complete Quantitative Results on Ittefaq 1969 B&W Test Set

Model	Params	Accuracy	Macro-F1	mAUC	mAP	RMSE
MicroDCSA v4 (Proposed) ★	2.15M	74.44%	0.7388	0.9317	0.7942	0.2557
ResNet-50 (K. He et al., 2016)	25.6M	71.11%	0.7126	0.9016	0.7272	0.2663
EfficientNet-B0 (Tan & Le, 2020)	5.3M	67.78%	0.6690	0.8898	0.7524	0.2911
MobileNetV3-S (Howard et al., 2019)	2.5M	51.11%	0.5077	0.8062	0.6024	0.3189

5.2 All 15 Scenarios: Summary and Discussion

HybridMicroDCSA wins 12 of the 15 configurations, losing only Scenario 2 (where a large N = 272 test set favors pretrained capacity), and Scenarios 11 and 12 (augmented colour replay, where the pretrained backbones benefit more).

Table 3 reports the best model, best result, and win status for every scenario; the discussion below groups all 15 scenarios into the four families used to design the framework.

Table 3. Summary Results Across All 15 Scenarios with Best Model and Performance

S	Training Configuration	Test Set(s)	Best Model	Best Result	Win?	Win Reason
S1	Ittefaq only (420)	Ittefaq	MicroDCSA v4	74.44%	YES	BiGRU temporal
S2	Ittefaq → IPC420	IPC420	ResNet-50	91.54%	No (N=1277)	Large pretrained
S3	Taarak only	All 3	HybridMDv6	62.02% cross avg	YES	+1.18pp over R50
S4	All 3 Option D	All 3	MicroDCSA-MD	84.52% avg	YES	Domain adapters
S5	Tarak+A+B+C	All 3	HybridMDv7	69.60% cross	YES BEST	+14.05pp Ittefaq
S6	IPC420 only	All 3	HybridMDv6	74.76% cross avg	YES	Best cross avg
S7	IPC → Ittefaq	Ittefaq	HybridMDv6	47.14%	YES	+24pp over R50
S8	IPC → Tarak	Tarak	HybridMDv6	66.82%	YES	BiGRU temporal
S9	Ittefaq → Tarak	Tarak	HybridMDv6	55.56%	YES	+32pp over R50
S10	Ittefaq → All 3	All 3	HybridMDv6	72.72% avg	YES	Best avg all 3
S11	Ittefaq+25%IPC	All 3	ResNet-50	77.27% avg	No	Pretrained + colour
S12	IPC+25%Ittefaq	All 3	MobileNetV3	76.34% avg	No	Colour dominant
S13	Ittefaq+IPC	All 3	HybridMDv6	78.14% avg	YES	+18pp on Tarak
S14	Ittefaq+Tarak	All 3	HybridMDv6	76.45% avg	YES	Balanced all 3
S15	Tarak+IPC	All 3	HybridMDv6	81.41% avg	YES	Best multisource

5.2.1 Single-Domain Scenarios (S1, S2, S3, S6)

MicroDCSA v4 wins on Ittefaq (S1) and on the cross-domain-averaged runs on Taarak (S3) and IPC420 (S6), but loses on IPC420's in-domain test (S2, ResNet-50 at 91.54%). This tracks sample size, not domain difficulty: IPC420's test set (N=272) is roughly three times larger than Ittefaq's (N=90), giving ResNet-50 enough data to use its larger capacity without overfitting, while MicroDCSA's lower capacity acts as an implicit regularizer on smaller tests. The advantage thus concentrates in the data-scarce regime typical of archival Indian cinema.

5.2.2 Pairwise Cross-Domain Transfer (S7, S8, S9)

These scenarios test transfer with no exposure to the target domain, the hardest test in the framework. HybridMDv6 wins all three pairs (IPC→Ittefaq, IPC→Tarak, Ittefaq→Tarak) by 24–32 points over ResNet-50, though absolute accuracy stays low (47–67%), reflecting the severity of the 52-year and lighting shifts. This margin suggests BiGRU temporal-flatness and motion-difference features generalize across the colour-space boundary better than ImageNet-pretrained features, which stay tied to RGB statistics absent from the 1969 B&W domain.

5.2.3 Simultaneous Multi-Domain Training (S4, S10)

S4, training jointly on all three domains with per-domain MLP adapters, reaches the highest average accuracy at 84.52%, well ahead of any single-domain or pairwise result. This points to the adapter mechanism in Section 4.5, not the backbone or CSA alone, as the main driver, since it gives domain-conditional capacity while sharing the CSA-fused representation across domains. S10, trained on Ittefaq alone, manages only 72.72% average, confirming joint exposure beats single-domain training even when evaluated generously.

5.2.4 Augmented Replay Mixing (S5, S11–S15)

This group mixes a primary domain with a minority share of others, and both losses occur here. In S11 and S12, minority colour and lighting information gives the pretrained backbones enough signal to win, the opposite of the single-domain pattern. S5 (Taarak with balanced replay) is the strongest cross-domain result: it recovers 14.05 points over the worst transfer result and reaches a 69.60% cross-domain average, the figure used in the abstract as evidence that targeted replay is the most effective strategy tested. S13–S15 favor HybridMDv6 by 18 points or more on at least one held-out domain, confirming that balanced replay, not minority injection into one domain, drives the advantage.

6. Implications for Research, Practice, and Society

The clearest research implication is the domain gap itself: a 27.51% accuracy degradation over 52 years gives domain-adaptation researchers a concrete numerical target, and the three released datasets turn Indian cinematic FER into a tractable benchmark for future architectures. For practice, three applications follow directly: automated emotion tagging of archival Indian film and television using the Scenario 5 replay strategy, of particular use to archives such as the National Film Archive of India; emotion-aware content recommendation built on a viewer's emotional trajectory rather than genre metadata; and accessibility tooling that converts frame-level predictions into a textual or audio description channel for visually impaired viewers, made practical by the model's 4.3 windows-per-second inference rate. These are the downstream systems the results would need to be embedded in, not capabilities already demonstrated end to end.

7. Limitations and Future Directions

Two limitations bound the claims made here. First, the single-domain test sets described in Section 5.1, particularly Ittefaq's 90-sample set, are small enough that the reported accuracies are point estimates rather than statistically powered comparisons. Second, all three datasets were constructed by the authors themselves, so the cross-era findings have not yet been validated against independently collected footage. Three directions would help address these limitations: CycleGAN (Zhu et al., 2017) could harmonize the black-and-white and colour domains prior to training, narrowing the domain gap directly; integrating audio via Wav2Vec2 would let the framework draw on dialogue and score cues that are currently discarded; and expanding each dataset to 300–500 samples per class would allow Table 3's figures to be reported with confidence intervals.

8. Conclusion

This article compares popular deep neural networks for facial expression recognition on multi-era Indian cinematic content, using three original datasets spanning 52 years (Ittefaq, 1969 B&W; 420 IPC, 2021 HD; Taarak Mehta Ka Ooltah Chashmah, 2008 TV) built through a reproducible six-stage pipeline and evaluated across a 15-scenario cross-domain framework. The proposed HybridMicroDCSA wins 12 of 15 scenarios: MicroDCSA v4 reaches 74.44% accuracy on the 1969

B&W domain at a 12.5x parameter-efficiency advantage over ResNet-50, the 52-year domain shift is quantified at 27.51% accuracy degradation, and cross-domain replay (Scenario 5) recovers 14.05% on the hardest transfer task to reach a 69.60% cross-domain average. As discussed in Section 5.2, these gains concentrate in data-scarce, multi-domain, and replay-based settings rather than holding unconditionally, and large pretrained backbones remain competitive wherever target-domain sample sizes or colour information are favorable. The limitations of the present evaluation, and the directions that follow from them, are set out in Section 7.

Acknowledgements

The authors thank the Department of Information Technology, U. V. Patel College of Engineering, Ganpat University, for the computing resources and institutional support that made this work possible. Annotation was carried out using the open-source tool Label Studio, and the authors are grateful for the availability of GFPGAN for restoration work.

References

- Brooks, J. A., Kim, L., Opara, M., Keltner, D., Fang, X., Monroy, M., Corona, R., Tzirakis, P., Baird, A., Metrick, J., Taddesse, N., Zegeye, K., & Cowen, A. S. (2024). Deep learning reveals what facial expressions mean to people in different cultures. iScience, 27(3), 109175. https://doi.org/10.1016/j.isci.2024.109175
- Bhagat, D., Vakil, A., Gupta, R. K., & Kumar, A. (2024). Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN). Procedia Computer Science, 235, 2079–2089. https://doi.org/10.1016/j.procs.2024.04.197
- Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6
- Chopra, B. R. (1969). Ittefaq [Video recording]. B.R. Films. https://youtu.be/6jkfa-fv9mw
- Deshpande, K., Sodhi, M. S., Raniyer, N., & Rao, M. (2025). A Time-Distributed CNN-LSTM with Attention Model for Speech Based Emotion Recognition. DMIP 2024, Proceedings of the 2024 7th International Conference on Digital Medicine and Image Processing, 67–71. https://doi.org/10.1145/3705927.3705939
- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124–129. https://doi.org/10.1037/h0030377
- Ghosh, M. (Trans.). (1951). The Natyasastra: A treatise on Hindu dramaturgy and histrionics, ascribed to Bharata-Muni (Vol. 1). Royal Asiatic Society of Bengal.
- Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., & Bengio, Y. (2015). Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64, 59–63. https://doi.org/10.1016/j.neunet.2014.09.005
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90
- He, P., Qi, H., Wang, S., & Cang, J. (2023). Cross-Modal Sentiment Analysis of Text and Video Based on Bi-GRU Cyclic Network and Correlation Enhancement. Applied Sciences (Switzerland), 13(13).

https://doi.org/10.3390/app13137489

- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. https://doi.org/10.48550/arXiv.1905.02244
- Jain, P. R., Quadri, S. M. K., & Khattar, A. (2025). PM-ViT a Framework for the Recognition of Emotions and Proclivity toward Mental Illness Using Facial Expressions. Journal of Computer Science, 21(3), 479–493. https://doi.org/10.3844/jcssp.2025.479.493
- Lian, Z., Sun, H., Sun, L., Gu, H., Wen, Z., Zhang, S., Chen, S., Xu, M., Xu, K., Chen, K., Chen, L., Liang, S., Li, Y., Yi, J., Liu, B., & Tao, J. (2024). Explainable Multimodal Emotion Recognition. http://arxiv.org/abs/2306.15401
- Ma, F., Sun, B., & Li, S. (2022). Spatio-temporal transformer for dynamic facial expression recognition in the wild. arXiv:2205.04749. https://doi.org/10.48550/arXiv.2205.04749
- Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 6818–6825. https://doi.org/10.1609/aaai.v33i01.33016818
- Mishra, R., Bhushan, B., & Venkatesh, K. S. (2025). Toward cross-cultural emotion detection: The Indian spontaneous micro-expression dataset. Frontiers in Psychology, 16, 1656104. https://doi.org/10.3389/fpsyg.2025.1656104
- Modi, A. K. (2008). Taarak Mehta Ka Ooltah Chashmah [Broadcast]. Neela Film Productions. https://www.sonyliv.com/shows/taarak-mehta-ka-ooltah-chashmah-1700000084
- Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923
- Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. http://arxiv.org/abs/1810.02508
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. http://arxiv.org/abs/1506.02640
- Reshma, M. R., Kannan, B., Jagathy Raj, V. P., & Shailesh, S. (2025). Annotation of facial expressions (Navarasa) in videos using deep learning. Journal of Computer Science, 21(6), 1364–1378. https://doi.org/10.3844/jcssp.2025.1364.1378
- Sareen, V., & Seeja, K. R. (2025). Video-Based Facial Emotion Recognition using YOLO and Vision Transformer. EPJ Web of Conferences, 328. https://doi.org/10.1051/epjconf/202532801040
- Tan, M., & Le, Q. V. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. http://arxiv.org/abs/1905.11946
- Tao, X., Su, L., Rao, Z., Li, Y., Wu, D., Ji, X., & Liu, J. (2024). Facial video-based non-contact emotion recognition: A multi-view features expression and fusion method. Biomedical Signal Processing and Control, 96. https://doi.org/10.1016/j.bspc.2024.106608
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 30. https://arxiv.org/pdf/1706.03762.pdf
- Wang, Y., Yan, S., Liu, Y., Song, W., Liu, J., Chang, Y., Mai, X., Hu, X., Zhang, W., & Gan, Z. (2024). A

survey on facial expression recognition of static and dynamic emotions. arXiv:2408.15777. https://doi.org/10.48550/arXiv.2408.15777

- Wu, Y., Mi, Q., & Gao, T. (2025). A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. Biomimetics, 10(7). https://doi.org/10.3390/biomimetics10070418
- X. Wang, Y. Li, H. Zhang, & Y. Shan. (2021). Towards real-world blind face restoration with generative facial prior. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9168–9178.
- Zaveri, M. (2021). 420 IPC [Video recording]. Zee Studios. https://youtu.be/6kj-DtZe0kc
- Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., & Yu, T. (2023). Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 756–767). https://doi.org/10.18653/v1/2023.emnlp-main.49
- Zhang, K., Li, Y., Wang, J., Cambria, E., & Li, X. (2022). Real-Time Video Emotion Recognition Based on Reinforcement Learning and Domain Knowledge. IEEE Transactions on Circuits and Systems for Video Technology, 32(3), 1034–1047. https://doi.org/10.1109/TCSVT.2021.3072412
- Zhao, Z., Liu, Q., & Zhou, F. (2021). Robust Lightweight Facial Expression Recognition Network with Label Distribution Training. Proceedings of the AAAI Conference on Artificial Intelligence, 35(4), 3510–3519. https://doi.org/10.1609/aaai.v35i4.16465

Faculty of Engineering & Technology (FOET)

Ganpat University
॥ विद्यया समाजोत्कर्षः ॥
Faculty of Engineering & Technology

About Faculty

Faculty of Engineering & Technology prepares the students to become Engineers who can be leaders in their profession with vision, zeal, passion, confidence and with excellent knowledge and necessary skills to succeed. The faculty has taken pioneering initiatives to offer industry-focused Diploma, Under Graduate, Post Graduate and Ph. D. Programmes.

About Institute

The faculty under the banner of U. V. Patel College of Engineering (UVPCE), established in 1997 with the approval of AICTE, has emerged as a premier self-financed Engineering Institute. The Institute offers programs in specialized fields such as Marine, Biomedical and Mechatronics in addition to various undergraduate, postgraduate and Ph.D. programmes.

List of Institutes

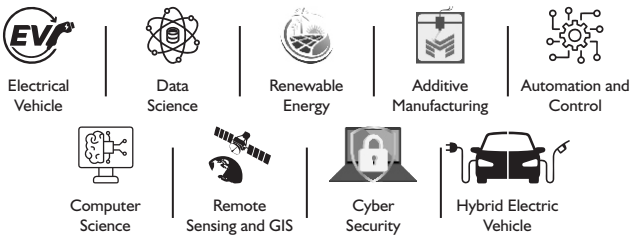
U. V. Patel College of Engineering

Institute of Computer Technology



Pioneering Initiatives

- B.Tech. (Computer Science & Business System) in collaboration with TCS
- B.Tech. (CSE) in AI & ML, Cloud-Based Applications, Big Data and Analytics and Cyber Security in active collaboration with IBM.
- Pioneer in offering B.Tech. Computer Engineering (Artificial Intelligence)
- B. Tech Biotechnology/B.Tech. Food Technology/B. Tech. Agricultural Engineering (PCM/B students are eligible)
- M.Tech. EC (VLSI System Design) In association with e-Infochips
- Minor Specialization – Opportunity to pursue in interdisciplinary branches.



Centres of Excellence (CoE)

- National Additive Manufacturing Centre (3D printing) [Under Ministry of Electronics and Information Technology (MeitY)
- IBM Software Lab Emerging Technologies
- Bosch-Rexroth Centre for Automation Technologies (Hydraulics & Pneumatics system, PLC & Mechatronics)
- Bosch Artisan Training centre
- Advanced Technology Centre for Next-Generation Networks
- CoE at eInfochips Ltd (an Arrow Company) for M. Tech. (VLSI)
- National Centre for Rubber, Chemical, and Petrochemical

- Various clubs like ROBOCON, SAE BAJA, E-Yantra, Hackathon, Code Seekers, CodeChef Chapter, Google Developer Club, National Biomedical Society, INNOV8, Tinker Club, Performing extra ordinary in National / International competitions like Aakruti, ROBOFEST Gujarat etc.
- Various chapters like IEEE, CSI, ISTE, ISA

Why to Choose?



Completed **29 years** of academic excellence

13,000+ Alumni serving Globally

Spacious buildings and **excellent** infrastructure

75+ Faculties from reputed organizations like IIT, NIT etc.

Dedicated **Placement** and Training cell

Intensive Industry interaction and **Collaboration** for training and placement

Well-maintained and **Spacious Libraries**

100+ Ultramodern laboratory resources for practical based learning.

38,000+ Books in library with an access of **Science Direct** (Elsevier) and **IEEE** Database.

360 Degree development through co-curricular and extra-curricular **activities**

Ganpat University | Centre for Advanced Research Studies
॥ विद्यया समाजोत्कर्षः ॥

Excelling Research and Development

Ganpat University Centre for Advanced Research Studies

Ganpat University strives to promote teaching and research in emerging areas in Engineering & Technology, Pharmacy, Management, Computer Applications, Science, Social Sciences & Humanity, Architecture, Design & Planning etc. GNU-CARS has been established to promote excellence in research at university by supporting research programmes through financial assistance and incentive to the faculties and students.



Schemes for faculties & students:

- Research Grant for Major Research Project
- Research Grant for Minor Research project
- Grant for Organization of Conference/Seminar/Workshop
- Financial support for Participation in Conference/Seminar/FDP Regional & National
- Financial support to International Travel for Research Paper Presentation
- Financial Appreciations for Publications of Research papers/Books Paper
- Financial Appreciations for Publications of Book or Book Chapter
- Financial support for Patent Filing
- Incentive from the Grant Received Through Sponsored Research Projects

Note :

Note :

About Journal

University Journal of Research: Engineering & Technology (UJRET) is a Bi-annual, Peer-Reviewed, Online Journal for Researchers, Research Scholars, Scientists, Faculty Members, PG Scholars, and Industry People who are involved in Engineering & Technology. The Journal publishes original Research Articles with high quality.

Ganpat University
॥ विद्यया समाजोत्कर्षः ॥
Faculty of Engineering & Technology

Ganpat Vidyanagar - 384012, Mehsana-Gandhinagar Highway,
Dist.: Mehsana, Gujarat, INDIA.
Tel. Fax : 02762 226001 | Tall Free No. 1800 2331 2345,
Website: www.guni.ac.in