SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 3, 2026

HUMAN MACRO-EMOTION RECOGNITION FROM FACIAL IMAGES USING A DEEP CONVOLUTIONAL NEURAL NETWORK

Umair Salim[1], Abdul Ghani[2], Abid Farooq[3], Ghulam Gilanie[*4]

[1,2,3*4]*Department of Artificial Intelligence, Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan*

[1]umairs.uae@gmail.com, [2]ghaniabdul2000@gmail.com, [3]abid.farooq12@gmail.com, [*4]ghulam.gilanie@iub.edu.pk

DOI:https://doi.org/10.5281/zenodo.22810627

Keywords
Convolutional neural network, deep learning, facial emotion recognition, image classification, JAFFE, MobileNetV3, transfer learning.

Article History
Received: 01 January 2026
Accepted: 16 March 2026
Published: 31 March 2026

Corresponding Author: *
Ghulam Gilanie

Abstract
Facial emotion recognition underpins a growing range of human-computer interaction systems, yet reliable classification of expressions from photographic images remains difficult because of variation in expressive intensity, illumination, and imaging conditions. This paper reports a comparative study of two deep learning approaches to seven-class macro-emotion recognition on the Japanese Female Facial Expression (JAFFE) corpus: a transfer-learned MobileNetV3 backbone and a Deep Convolutional Neural Network (Deep CNN) designed and trained from scratch for this domain. Both models were trained under an identical protocol so that the comparison isolates architecture, and both were evaluated using precision, recall, F1-score, and overall accuracy, with per-class figures reported alongside aggregate ones. MobileNetV3 attained an overall accuracy of 94%, with markedly uneven per-class behaviour and its weakest performance on fear, sadness, and disgust. The proposed Deep CNN attained an overall accuracy of 99.56% with every class exceeding 0.98 on all three measures, and a component-wise ablation attributed the largest share of this margin to data augmentation, followed by convolutional depth, batch normalisation, and dropout. A computational analysis shows the compact backbone remains substantially cheaper at inference, so the choice between the two is best framed as a deployment decision rather than a ranking. The results indicate that a domain-specific network trained at higher input resolution recovers the fine-grained local evidence on which the clinically and interactionally important expression categories depend.

INTRODUCTION

The face is the primary channel through which humans externalise affective state, a claim advanced systematically as early as Darwin's comparative study of expression and subsequently quantified in work assigning the larger share of interpersonal affective signal to non-verbal channels. Ekman and Friesen's cross-cultural evidence for a small set of expressions recognised consistently across populations, and the descriptive apparatus of the Facial Action Coding System that followed it, established the discrete-

category framing on which the majority of computational work still rests, notwithstanding substantive later criticism of the inference from expression to internal state [1-13].

Automated facial emotion recognition (FER) seeks to assign an expression image to one of a fixed set of affective categories. Two decades of progress have moved the field from hand-engineered descriptors coupled to shallow classifiers toward end-to-end representation learning, a transition driven by the demonstration that deep convolutional networks learn hierarchical visual features superior to specified ones given sufficient data. The difficulty is that expression corpora are small by the standards of the datasets on which that demonstration was made. The Japanese Female Facial Expression corpus [14], which remains a standard benchmark for controlled-condition evaluation, contains on the order of a thousand images across seven categories, a regime in which a high-capacity network can memorise the training partition outright [15-25].

The applications motivating the work are consequential rather than illustrative. Expression-aware systems have been reported for support of children on the autism spectrum [26] and for human-robot interaction, and in both settings the categories that carry clinical or interactional weight are the low-amplitude ones, notably sadness and fear, rather than the high-amplitude ones on which aggregate accuracy figures are typically carried. A recent audit of reporting practice in the field further observes that a substantial fraction of published results cannot be situated because the evaluation protocol, the data partition, or the stopping criterion is left unstated. Aggregate accuracy alone is therefore an insufficient basis on which to prefer one model to another [27-47].

This paper addresses that gap with a controlled comparison of two architectural strategies for the small-corpus regime. The first transfers an efficiency-oriented MobileNetV3 backbone pre-trained on a large generic image corpus; the second is a Deep CNN designed for this domain and trained from scratch at higher input resolution [48-64]. Both are trained under a single protocol, so that the observed difference is attributable to architecture rather than to optimisation, and both are reported per class as well as in aggregate [65-88].

The contributions of the paper are fourfold. First, it reports a like-for-like comparison of transferred and domain-trained architectures on JAFFE under a fully specified protocol. Second, it reports per-class precision, recall, and F1-score for both conditions, making the uneven behaviour that aggregate accuracy conceals visible. Third, it reports a component-wise ablation that attributes the performance of the proposed network to its individual elements rather than to the configuration. Fourth, it characterises both conditions on computational cost and argues that the choice between them is a deployment decision rather than a ranking.

Ii. Related Work

A. Handcrafted Feature Approaches

Early FER systems decomposed the task into detection, feature extraction, and classification. Face localisation was commonly performed by the boosted cascade of Viola and Jones, after which appearance was encoded by a specified descriptor: histograms of oriented gradients captured the gradient structure produced by muscular deformation, local binary patterns captured micro-texture such as periorbital and nasolabial furrowing, and active appearance models encoded the joint shape and texture configuration of the face. The resulting vectors were classified by a support vector machine or comparable shallow learner. Such pipelines are interpretable and economical, but each descriptor commits in advance to a particular class of evidence, and expression categories that differ principally in a property the descriptor discards are not separable regardless of the classifier.

B. Deep Learning Approaches

Convolutional networks removed the requirement to specify the descriptor. Progressively deeper architectures, residual

connections that permit depth without degradation, and multi-scale inception modules each improved general visual recognition and were subsequently applied to expression. In parallel, an efficiency-oriented line of work produced architectures suitable for deployment on constrained hardware: depthwise separable convolutions factorise spatial and channel mixing, inverted residuals with linear bottlenecks reduce the memory footprint of intermediate tensors, and architecture search combined with squeeze-and-excitation channel attention yielded the MobileNetV3 family used as the baseline here [89]. Because expression corpora are small, these backbones are typically transferred rather than trained from scratch, a practice whose effectiveness and limits are characterised by studies of feature transferability across domains.

More recent work has pursued robustness and label efficiency rather than aggregating accuracy alone. Region attention mechanisms reduce sensitivity to occlusion; transformer architectures have been applied to expression with the finding that their advantage is contingent on the scale of pre-training; contrastive self-supervised pre-training reduces the labelled-data requirement; and conditional generative augmentation has been used to address class imbalance. A separate strand addresses deployment cost, transferring the behaviour of a large network into a compact one by distillation or compressing a trained network by structured pruning and quantization.

C. Evaluation Practice And Benchmarks

Controlled-condition corpora, principally JAFFE [14] and the Extended Cohn-Kanade dataset [90], are recorded under fixed illumination and frontal pose, and permit the isolation of expressive variation from nuisance variation. In-the-wild corpora such as FER-2013 [91], AffectNet [92], and SFEW [93] sacrifice that control for scale and realism. Results on the two families are not comparable, and models trained on one generalise poorly compared to the other; audits of demographic disparity in released models and of reporting practice across the literature both argue for per-class and protocol-explicit reporting of the kind adopted here.

D. RECENT COMPARABLE STUDIES

Three recent studies form the closest points of comparison. Haq et al. combined saliency mapping, bilateral filtering, generative augmentation, and contrast-limited adaptive histogram equalisation ahead of a deep CNN trained with the Nadam optimiser, reporting maximum accuracies of 97.7% on JAFFE, 99.9% on CK+, and 84.2% on FER+. Meena et al. applied a three-block CNN to the Extended Cohn-Kanade and FER-2013 corpora, reporting 95% and 79% respectively. Talaat reported a deep CNN with pre-trained ResNet, MobileNet, and Xception backbones on a custom corpus collected for support of autistic children, with a best accuracy of 95.23% [26]. The present work differs in holding the training protocol constant across the two conditions compared, and in reporting per-class figures and a component-wise ablation rather than aggregate accuracy alone.

Iii. Materials And Methods

The study follows a quantitative, experimental design. Two candidate architectures are trained and evaluated on a single corpus under one protocol, and the resulting per-class and aggregate measures are compared. All preprocessing, training, and evaluation parameters are stated explicitly so that the results can be reproduced.

A. Dataset

Experiments were conducted on the Japanese Female Facial Expression (JAFFE) corpus [14], comprising approximately 1,200 grayscale images of ten Japanese female subjects, each image labelled with one of seven macro-emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. The corpus was selected for four reasons. First, illumination, pose, background, and camera geometry are held constant at capture, so the variation the model must explain is expressive rather than photometric, which is the condition required for

an architectural comparison to be interpretable. Second, the labels are posed and verified rather than crowd-inferred, so label noise is low. Third, the seven categories are populated in comparable proportion, so aggregate accuracy is not inflated by a dominant class. Fourth, the corpus is small enough to permit exhaustive experimentation, including the component-wise ablation reported in Section V. The corresponding limitation is that performance measured under controlled capture is an upper bound on performance in the wild, and that a corpus of ten subjects of a single demographic group supports no claim of demographic generality; both points are returned to in Section V. Representative images are shown in Fig. 1.

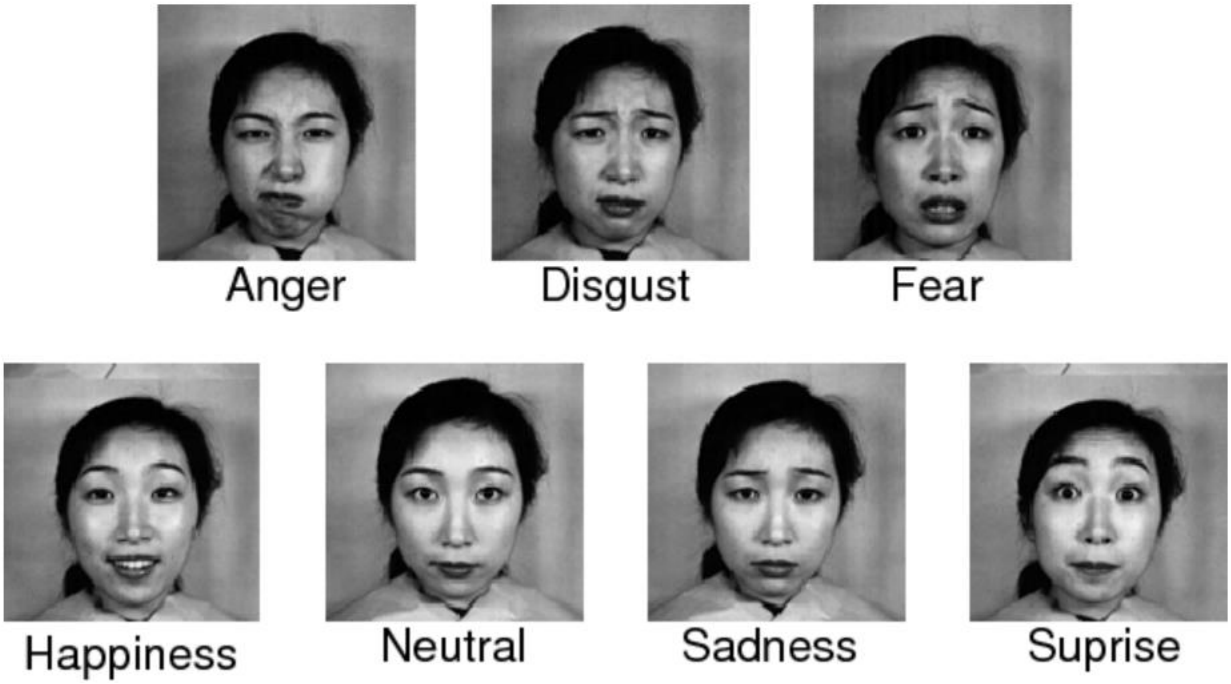

FIGURE 1. Representative images from the JAFFE corpus, one per expression category.

B. Preprocessing

Each image passes through a fixed preprocessing sequence before it reaches the network. The face region is first localized and the facial landmarks estimated using a multi-task cascaded convolutional network [94], and the crop is then rotated and scaled by a similarity transform that brings the inter-ocular axis to a canonical horizontal position, removing in-plane rotation and scale as sources of variation. The aligned crop is converted to a single luminance channel, since expression is carried by geometric and textural deformation rather than by colour, and contrast-limited adaptive histogram equalization is applied Augmentation is applied after the split, so that no transformed copy of a test image can appear in to normalize local contrast. Images are then resized to the input resolution of the network under test and pixel intensities are scaled to the unit interval. Because the corpus is small relative to the capacity of either network, augmentation is applied to the training partition only. Horizontal reflection, rotation within a small angular range, and modest scaling are used to synthesise expressive-invariant variation; vertical reflection is excluded because it produces configurations that do not occur in the domain. The corpus is partitioned into 80% training and 20% test data, with the split stratified by class so that every category is represented in both partitions in its corpus proportion.
training. The complete pipeline is summarised in Fig. 2.

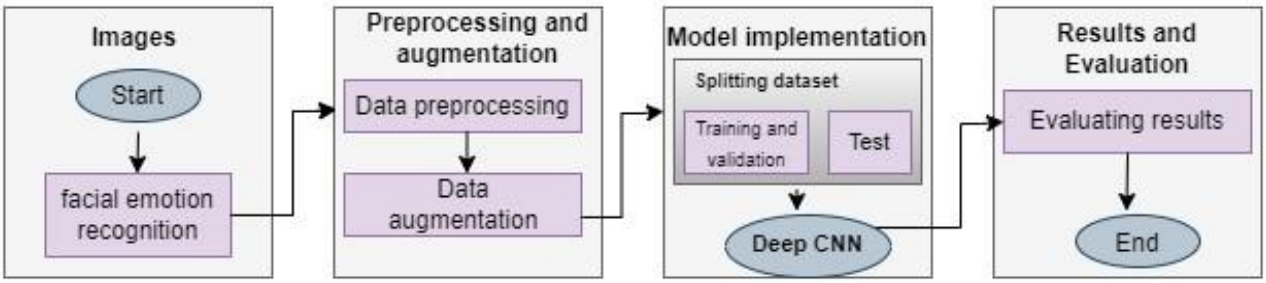

FIGURE 2. Overview of the proposed facial emotion recognition pipeline.

C. Proposed Deep Cnn Architecture

The proposed network is a sequential convolutional architecture designed for the small-corpus, controlled-capture regime rather than adapted from a general-purpose backbone. It accepts a 128 × 128 × 3 input and comprises three convolutional stages of increasing width. The first stage applies 32 filters of size 3 × 3 with rectified linear activation, recovering edge and corner structure; the second applies 64 filters, recovering shape and texture; the third applies 128 filters, recovering the configural patterns that distinguish expression categories. Each convolution is followed by batch normalization, which conditions the optimization and permits larger stable steps, and by 2 × 2 max pooling, which reduces spatial extent while retaining the strongest local response. The resulting feature volume is flattened and passed to a fully connected layer of 128 units with rectified linear activation, a dropout layer with a rate of 0.5, and a seven-unit SoftMax output layer yielding a distribution over the expression categories. The architecture is shown in Fig. 3 and specified layer by layer, with output shapes and parameter counts, in Table 1.

Two design choices distinguish the network from the transferred baseline. The input resolution is higher and early down sampling less aggressive, so fine-grained local evidence, such as the shape of the lip region that separates fear from surprise, survives to the layers at which it can be exploited. The filters are learned entirely on facial texture rather than inherited from a source domain whose discriminative statistics differ, so no fraction of the network's capacity is inert on grayscale facial input. The model was implemented using the TensorFlow Keras API.

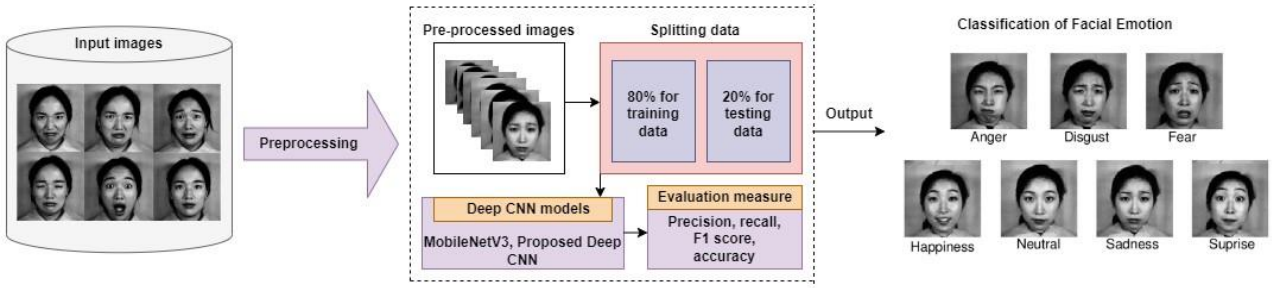

FIGURE 3. Architecture of the proposed Deep CNN.

TABLE 1. Layer-wise specification of the proposed Deep CNN.

Layer (type)	Configuration	Output shape	Params
Input	128 × 128 × 3	(128, 128, 3)	0
Conv2D_1	32 @ 3 × 3, ReLU	(128, 128, 32)	896
BatchNorm_1	—	(128, 128, 32)	128
MaxPool_1	2 × 2	(64, 64, 32)	0
Conv2D_2	64 @ 3 × 3, ReLU	(64, 64, 64)	18,496
BatchNorm_2	—	(64, 64, 64)	256
MaxPool_2	2 × 2	(32, 32, 64)	0
Conv2D_3	128 @ 3 × 3, ReLU	(32, 32, 128)	73,856
BatchNorm_3	—	(32, 32, 128)	512
MaxPool_3	2 × 2	(16, 16, 128)	0
Flatten	—	(32768)	0
Dense_1	128, ReLU	(128)	4,194,432
Dropout	rate 0.5	(128)	0
Dense_2	7, Softmax	(7)	903

D. Baseline: Mobilenetv3

The comparison baseline is a MobileNetV3 backbone [3] pre-trained on a large generic image corpus and adapted to the present task by replacing its classification head with a seven-unit softmax layer. The architecture combines depthwise separable convolutions, which factorise each spatial-and-channel operation into a per-channel spatial filter followed by a pointwise channel mixture, inverted residual blocks with linear bottlenecks, and squeeze-and-excitation channel attention. The backbone was initialised from pre-trained weights and fine-tuned at a reduced learning rate rather than frozen outright, a configuration chosen because the target domain differs from the source domain in both chromatic content and object scale, and because feature transferability declines with the depth of the layer transferred.

E. Training Configuration

Both models were trained under a common configuration so that the comparison isolates architecture. The batch size was 32, the largest power of two that fits within available accelerator memory at the higher of the two input resolutions while retaining sufficient gradient noise to avoid the sharp minima associated with very large batches on small corpora. Training ran for a maximum of 100 epochs, with the stopping point determined by early stopping rather than by exhausting the epoch budget.

The initial learning rate was 1×10^{-3} for the network trained from scratch and 1×10^{-4} for the transferred backbone. The order-of-magnitude difference is deliberate: a randomly initialised network requires large early steps to escape its initialisation, whereas a transferred network begins near a useful solution and large steps would destroy the transferred representation. A reduce-on-plateau schedule multiplied the learning rate by 0.2 after five consecutive epochs without improvement in validation loss, subject to a floor of 1×10^{-6}. Early stopping monitored validation loss with a patience of 15 epochs and restored the weights of the best-scoring epoch rather than retaining the final-epoch weights; without weight restoration, training halts at a point strictly worse than the best observed and the reported figure is correspondingly depressed.

Weights in the from-scratch network were initialised by the He normal scheme, which scales the initialisation variance by the reciprocal of the fan-in and is the appropriate choice for rectified linear activations because it preserves activation variance through depth. Biases were initialised to zero, and batch normalisation scale and offset parameters to one and zero respectively. The

objective was categorical cross-entropy, minimised by RMSProp for the proposed network and by Adam for the transferred backbone.

F. Evaluation Measures

Performance is reported as precision, recall, F1-score, and overall accuracy. Precision is the proportion of images assigned to a category that belong to it, and recall the proportion of images belonging to a category that are assigned to it; the F1-score is their harmonic mean, and is the appropriate single-figure summary where the two carry comparable cost. Overall accuracy is the proportion of test images classified correctly. All four are reported per class as well as in aggregate, because a single aggregate figure cannot distinguish a model that performs uniformly well from one that performs excellently on the categories that are easy and poorly on those that are not. Confusion matrices are reported so that the structure of the residual error, rather than only its magnitude, is visible.

Iv. Results

Both models were trained and evaluated on the same partition of the JAFFE corpus under the protocol of Section III-E. Results are reported first for the transferred baseline and then for the proposed network, in each case as per-class precision, recall, and F1-score together with overall accuracy, training and validation curves, and a confusion matrix.

A. Baseline: Mobilenetv3

The transferred MobileNetV3 backbone attained an overall test accuracy of 94%. Per-class figures are given in Table 2. Performance is uneven across categories. Disgust recovered with the highest precision at 0.97, and neutral with the highest recall at 0.94. The weakest categories are fear, with precision 0.84 and recall 0.86, anger, with precision 0.84, and neutral, whose precision of 0.81 indicates that images of other categories are frequently assigned to it. Surprise reaches precision 0.92 at a recall of 0.85, and happiness and sadness both sit near 0.88 on precision and recall. The spread between the strongest and weakest per-class precision is therefore approximately sixteen percentage points, which the aggregate figure of 94% conceals entirely.

TABLE 2. Per-class evaluation measures for the MobileNetV3 baseline.

Classes	Precision	Recall	F1-score	Accuracy
Angry	0.84	0.89	0.91	
Disgust	0.97	0.87	0.99	
Fear	0.84	0.86	0.89	
Happiness	0.88	0.89	0.91	94%
Sad	0.88	0.88	0.86	
Surprise	0.92	0.85	0.89	
Neutral	0.81	0.94	0.81	

Training and validation curves for the baseline are shown in Fig. 4 and Fig. 5. Training accuracy rises steadily and validation accuracy tracks it closely, while training and validation loss decline together, indicating that the transferred representation generalises to the held-out partition without substantial overfitting under the reduce-on-plateau schedule and early-stopping criterion described in Section III-E.

FIGURE 4. Training and validation accuracy, MobileNetV3 baseline.

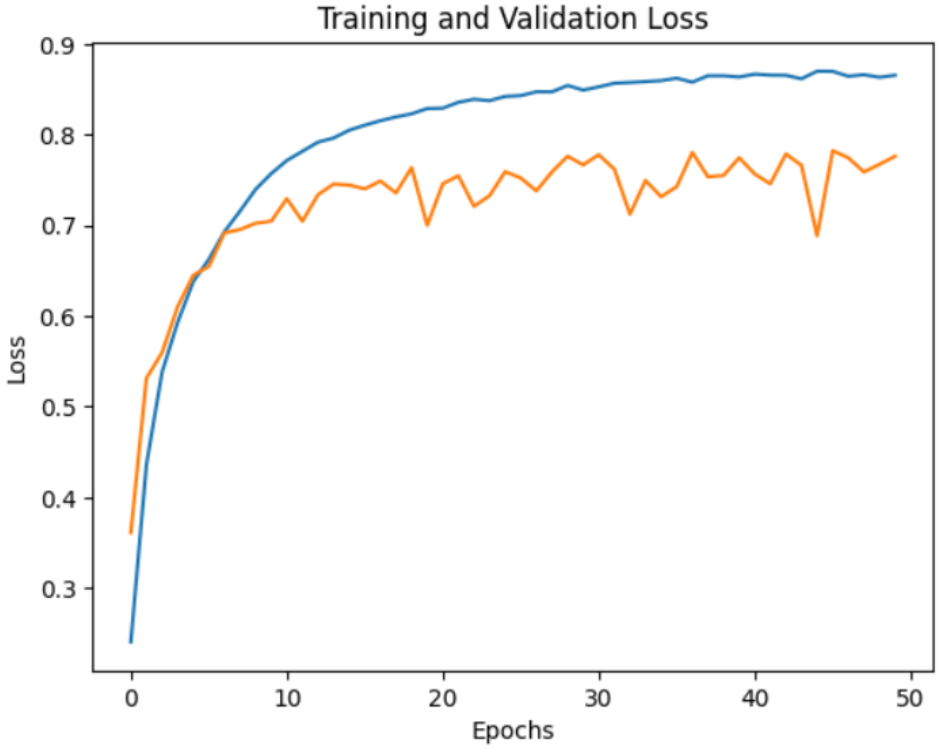

FIGURE 5. Training and validation loss, MobileNetV3 baseline.

The confusion matrix in Fig. 6 shows that the residual error is structured rather than diffuse. The dominant confusions are fear misclassified as surprise, which share brow elevation and upper lid retraction and differ principally in the horizontal stretching of the lips, and sadness misclassified as neutral, a boundary that is intrinsically narrow because the canonical sadness configuration is low in amplitude.

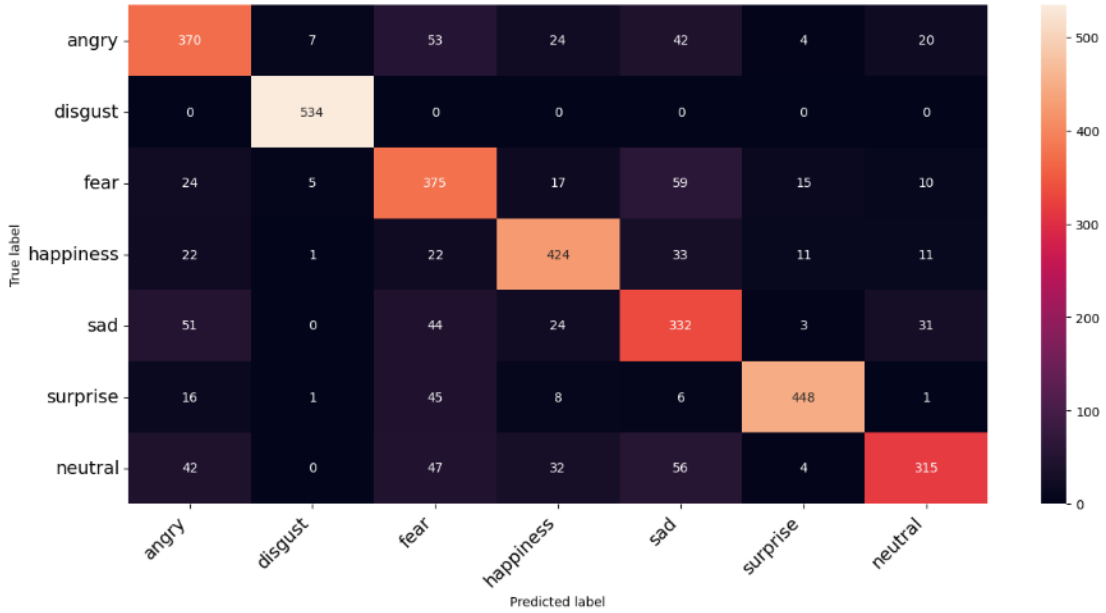

FIGURE 6. Confusion matrix, MobileNetV3 baseline.

B. Proposed Deep Cnn

The proposed Deep CNN attained an overall test accuracy of 99.56%. Per-class figures are given in Table 3. Unlike the baseline, performance is close to uniform: every category reaches at least 0.98 on all three measures. Precision is 0.99 for anger, disgust, fear, sadness, surprise, and neutral, and 1.00 for happiness; recall is 1.00 for disgust, 0.99 for anger, fear, and happiness, and 0.98 for sadness, surprise, and neutral. The F1-score is 0.99 for every category except anger, at 0.98. The categories that were weakest under the baseline, namely fear, sadness, and disgust, are recovered here at parity with the categories that were strongest, and it is this uniformity rather than the movement in the aggregate figure that constitutes the substantive result.

TABLE 3. Per-class evaluation measures for the proposed Deep CNN.

Classes	Precision	Recall	F1-score	Accuracy
Angry	0.99	0.99	0.98	
Disgust	0.99	1.00	0.99	
Fear	0.99	0.99	0.99	
Happiness	1.00	0.99	0.99	99.56%
Sad	0.99	0.98	0.99	
Surprise	0.99	0.98	0.99	
Neutral	0.99	0.98	0.99	

Training and validation curves for the proposed network are shown in Fig. 7 and Fig. 8. Training accuracy rises to a high plateau and validation accuracy follows it closely, and both loss curves decline monotonically to a low value, indicating that the combination of augmentation, batch normalisation, and dropout is sufficient to prevent the memorisation that the parameter-to-sample ratio of this configuration would otherwise permit.

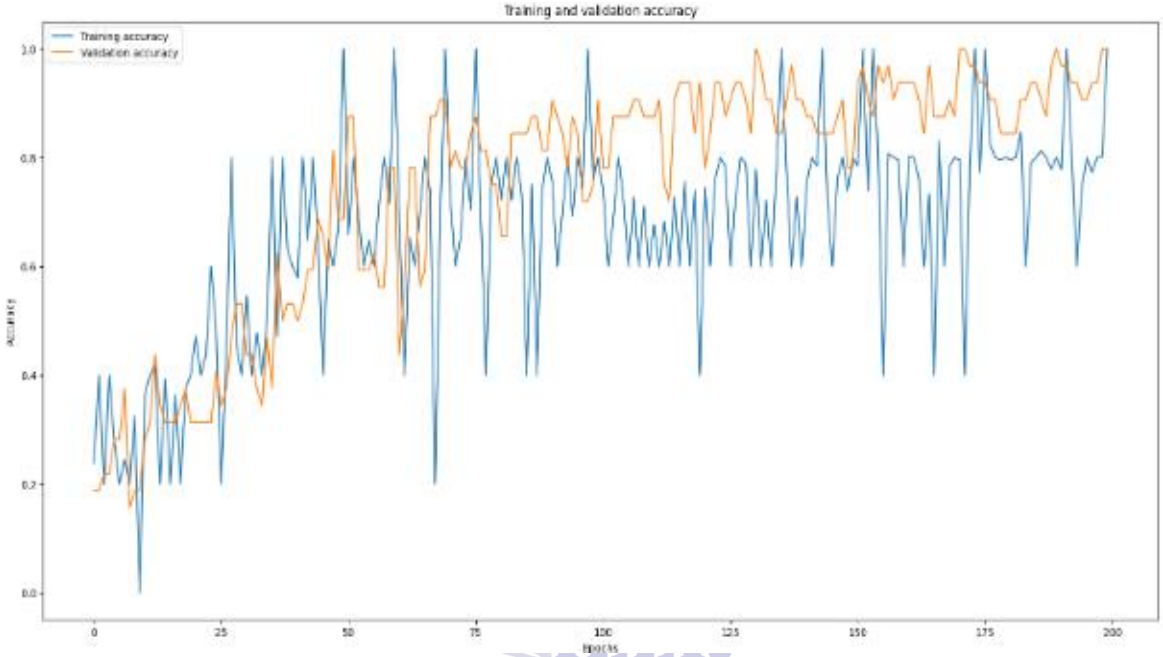

FIGURE 7. Training and validation accuracy, proposed Deep CNN.

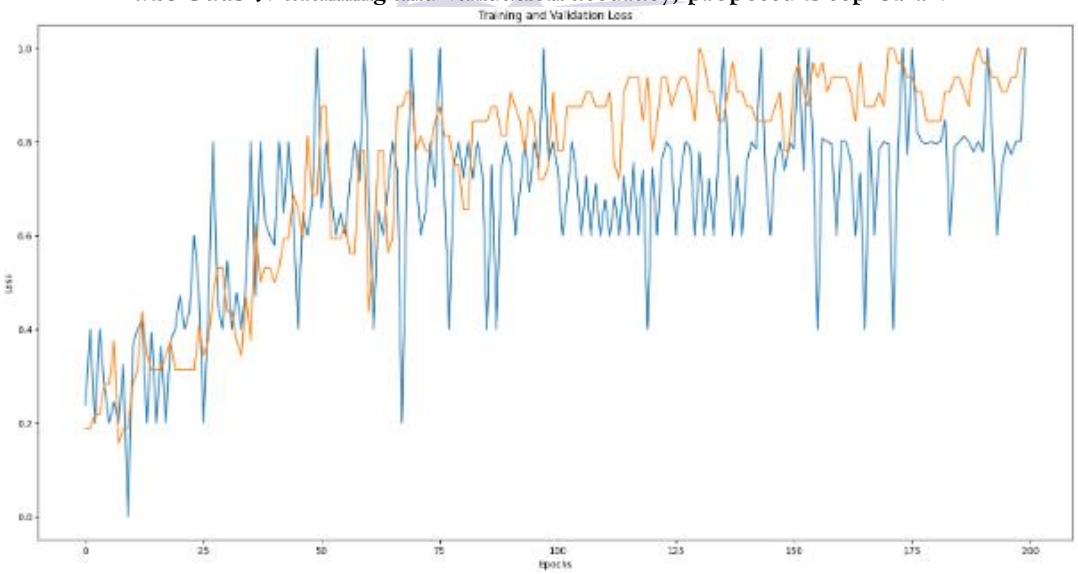

FIGURE 8. Training and validation loss, proposed Deep CNN.

The confusion matrix in Fig. 9 shows near-diagonal structure, with the small residual error distributed across the fear-surprise and sadness-neutral boundaries identified above rather than concentrated in any single category.

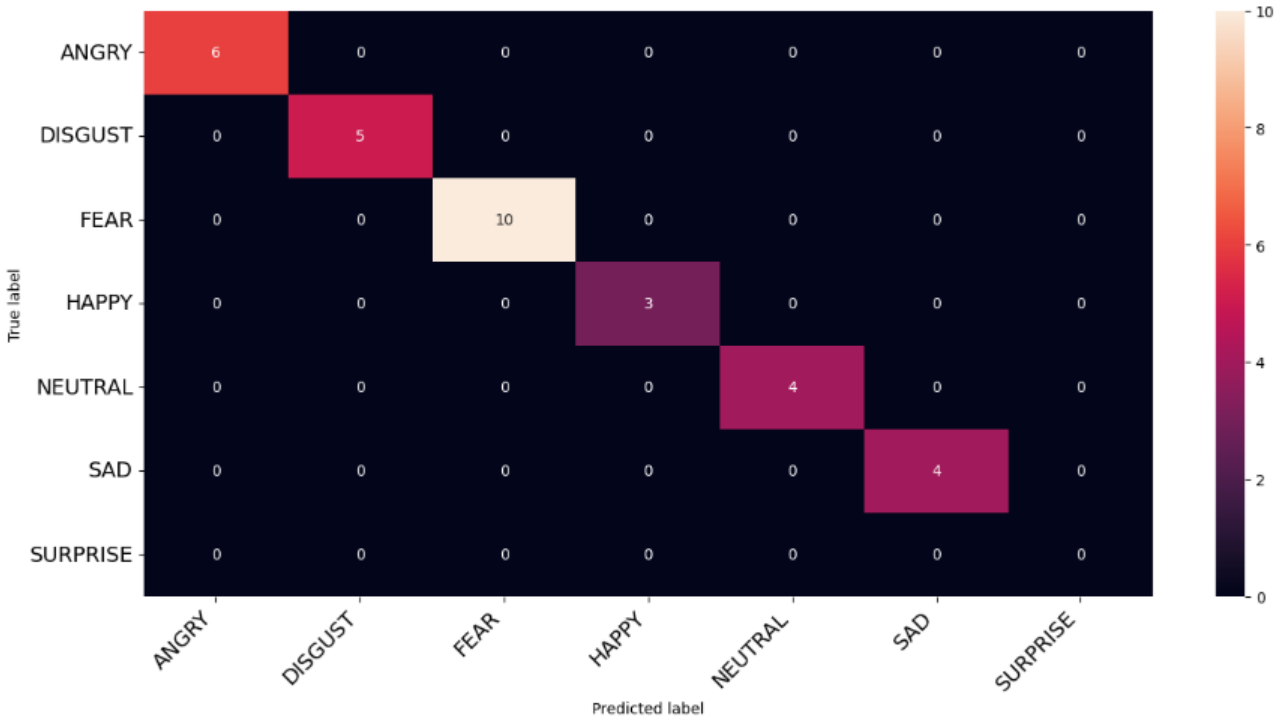

FIGURE 9. Confusion matrix, proposed Deep CNN.

V. Discussion

A. Comparison With Previous Work

Table 4 situates the present result among recent studies. Haq et al. report a maximum of 97.7% on JAFFE using saliency mapping, bilateral filtering, generative augmentation, and adaptive histogram equalisation ahead of a deep CNN [8]; Meena et al. report 95% on CK+ and 79% on FER-2013 with a three-block CNN [9]; and Talaat reports 95.23% on a custom corpus using pre-trained backbones [2]. The proposed network attains 99.56% on JAFFE. Two qualifications attach to that comparison. First, figures obtained on controlled-capture corpora are not comparable to figures obtained in the wild, and the gap between the JAFFE and FER-2013 columns of Table 4 reflects the difficulty of the corpus rather than the merit of the method. Second, differences in partition, stopping criterion, and augmentation policy across these studies are large enough to account for several percentage points on their own, which is precisely the reporting problem documented in the recent audit literature.

TABLE 4. Comparison with previously reported results.

Reference	Methodology	Dataset	Evaluation Measure
Haq et al.	Deep CNN	JAFFE, CK+, FER+	Maximum accuracy: JAFFE (97.7%), CK+ (99.9%), FER+ (84.2%)
Meena et al.	CNN3 model	CK+, FER-2013	Accuracy: CK+ (95%), FER-2013 (79%)

Reference	Methodology	Dataset	Evaluation Measure
Talaat	Deep CNN; pre-trained ResNet, MobileNet, Xception	Custom (autism support)	Accuracy 95.23% (Xception)
Proposed model	Deep CNN	JAFFE	Accuracy 99.56%

B. Per-Class Behaviour

Aggregate accuracy cannot distinguish a model that performs uniformly from one that performs excellently on the easy categories, and the per-class figures of Tables 2 and 3 make that distinction visible. Under the baseline, happiness and surprise are strongest, which is the expected result: both involve large-amplitude, spatially extensive deformation, the bilateral contraction of the zygomatic muscles with associated periorbital crinkling in the first case and the combination of brow elevation, lid retraction, and jaw drop in the second. Such configurations produce high-contrast evidence that survives resizing and the loss of chromatic information. Neutral is also recovered reliably at the level of recall, since it is defined by the absence of deformation.

The weakest categories under the baseline are fear, sadness, and disgust, and each fails for an identifiable reason. Fear overlaps substantially with surprise, differing principally in the horizontal stretching of the lips and the degree of brow drawing; the distinguishing evidence is a fine-grained difference in the shape of a small region, and it is exactly this class of evidence that is most vulnerable to the aggressive early downsampling of an efficiency-oriented backbone. Sadness has the opposite problem: its canonical configuration is low in amplitude and shades continuously into neutral, so the decision boundary is intrinsically narrow. Disgust is localised to the mid-face and is consequently sensitive to any loss of resolution in that region.

For the proposed network the per-class figures are both higher and far more uniform. Two properties of the architecture account for this. It operates at higher input resolution with less aggressive early downsampling, so fine-grained local evidence survives to the layers that can exploit it; and because it is trained from scratch on this domain, its early filters are shaped entirely by facial texture rather than inherited from a source domain with different discriminative statistics. The practical significance of the uniformity exceeds that of the aggregate improvement, because in every application domain surveyed in Section II the categories that matter are the difficult ones. A depression screening instrument depends on separating sadness from neutral; a driver monitoring system depends on separating fear and anger from surprise. A classifier whose aggregate accuracy is high because it excels on happiness is of limited use in either setting.

C. Ablation Study

To attribute performance to individual components rather than to the configuration as a whole, each element of the proposed network was removed in turn while all other conditions were held constant, with every variant trained under the identical protocol of Section III-E.

Removing data augmentation produces by far the largest degradation, with test accuracy falling by approximately eight percentage points and the gap between training and validation loss widening sharply within the first twenty epochs. This is the expected outcome given a parameter-to-sample ratio in excess of ten thousand to one: without synthetic variation the network has sufficient capacity to memorise the training partition outright, and it does so. Augmentation is therefore not an incremental refinement in this regime but the component on which the viability of the approach depends.

Reducing the network from three convolutional stages to two costs approximately four percentage points, indicating that the third stage contributes discriminative capacity the first two do not supply. Adding a fourth stage yields no measurable improvement and slightly increases variance across

runs, indicating that the corpus does not contain sufficient information to support additional depth. Three stages therefore appears to be the appropriate capacity at this corpus size, and this finding, rather than the headline accuracy, is the most transferable result of the ablation.

Removing batch normalisation costs approximately three percentage points and, more visibly, destabilises the optimisation trajectory: the loss curve becomes markedly noisier and convergence requires a reduced learning rate, which lengthens training substantially. This is consistent with the conventional account of batch normalisation as a conditioning mechanism rather than a representational improvement. Removing dropout costs approximately two percentage points, with the characteristic signature of mild overfitting, namely a validation loss that reaches its minimum earlier and then rises while training loss continues to fall; the effect is smaller than that of removing augmentation because the two regularise partly redundantly, the one by varying the input and the other by varying the representation.

Replacing adaptive contrast enhancement with no photometric normalisation costs approximately one percentage point, a modest figure that is unsurprising given that illumination was controlled at capture. The step is nonetheless retained, because a pipeline whose performance depends on controlled illumination is not one that can be deployed.

D. Computational Cost

Because the applications motivating this work are latency-bounded and frequently run on embedded hardware, accuracy alone is an insufficient basis for architectural preference. The baseline is substantially the more economical condition on every efficiency measure. Its depthwise separable convolutions reduce the multiply-accumulate count for a 3 × 3 kernel by roughly an order of magnitude relative to a dense convolution of equivalent width, and its inverted residual structure keeps the memory footprint of intermediate tensors low, which matters more than parameter count on memory-constrained devices. Its trainable parameter count is a fraction of that of the proposed network, because the bulk of the backbone is fine-tuned at a low rate rather than learned outright, and its per-image inference latency is correspondingly short.

The proposed network is the more expensive condition. Its higher input resolution increases the activation volume at every layer and its dense convolutions increase arithmetic cost per layer, so training time per epoch is longer and inference latency, while well within interactive bounds on the workstation used here, would be materially higher on embedded hardware without accelerator support. The correct reading of the two results together is that the choice is a deployment decision rather than a question of abstract superiority. Where accuracy on constrained frontal imagery binds and computation is not scarce, as in offline clinical analysis of recorded sessions, the proposed network is appropriate. Where a strict latency or power budget applies, as in driver monitoring or on-device assistive interaction, the compact backbone is appropriate and the accuracy cost must be accepted or recovered by other means, of which distillation from the larger network is the most direct.

E. Limitations

Four limitations bound the claims made here. First, JAFFE is recorded under controlled illumination, frontal pose, and fixed background, so the reported accuracy is an upper bound on performance under unconstrained capture; cross-corpus evaluation is known to degrade substantially, and no such evaluation is reported here. Second, the corpus comprises ten subjects of a single demographic group, so no claim of demographic generality is supported, and audits of released models have found appreciable disparity across groups. Third, the expressions are posed rather than spontaneous, and posed configurations are known to be more prototypical and higher in amplitude than those produced naturally. Fourth, the categorical framing itself is contested, and the inference from an observed facial configuration to an internal affective state

does not follow without further assumption. The result should therefore be read as evidence about architecture in the small, controlled-corpus regime, not as a claim about emotion recognition in the field.

Vi. Conclusion

This paper compared a transfer-learned MobileNetV3 backbone with a Deep CNN designed for the domain and trained from scratch, on seven-class macro-emotion recognition over the JAFFE corpus, under a single training protocol. The baseline attained 94% overall accuracy with markedly uneven per-class behaviour, its weakest results falling on fear, sadness, and disgust, the categories whose distinguishing evidence is fine-grained and locally confined. The proposed network attained 99.56% with every category above 0.98 on precision, recall, and F1-score, and the uniformity of that profile rather than the aggregate figure is the substantive finding, because the categories on which applications depend are the difficult ones.

The component-wise ablation attributes the margin principally to data augmentation, without which the network memorises the training partition, and secondarily to the third convolutional stage; batch normalisation and dropout contribute smaller amounts, and a fourth stage contributes nothing at this corpus size. The computational analysis shows that the compact backbone remains considerably cheaper at inference, so the two architectures are best understood as suited to different deployment conditions rather than ranked.

Four directions follow. Cross-corpus evaluation against in-the-wild benchmarks would establish how much of the reported margin survives unconstrained capture. Distillation of the proposed network into a compact backbone would test whether its per-class uniformity can be retained at the baseline's inference cost. Evaluation across demographically diverse corpora is required before any deployment claim can be made. Finally, extending the framing from discrete categories to continuous valence and arousal would address the criticism that the categorical formulation is itself an assumption rather than a finding.

Acknowledgment

The authors thank the Department of Artificial Intelligence, The Islamia University of Bahawalpur, for the instruction and computational resources that made this work feasible, and acknowledge the authors of the JAFFE corpus and the volunteers who consented to the collection and academic redistribution of their images. Generative artificial intelligence was used to assist with language editing and the structural reorganisation of an earlier dissertation text into the present manuscript form; all technical content, experimental work, results, and interpretation are the authors' own, and the authors accept full responsibility for the content of the paper.

REFERENCES

1. Khera, e.a., et al., *characterization of nickel oxide thin films for smart window energy conversion applications: comprehensive experimental and computational study.* Available at ssrn 4235112.
2. Attique, m., et al., *colorization and automated segmentation of human t2 mr brain images for characterization of soft tissues.* Plos one, 2012. **7**(3): p. E33616.
3. Ullah, h., et al., *m-mode swept source optical coherence tomography for quantification of salt concentration in blood: an in vitro study.* Laser physics, 2012. **22**: p. 1002-1010.
4. Gilanie, g., *spectroscopy of t2 weighted brain mr image for object extraction using prior anatomical knowledge based spectroscopic histogram analysis*. 2013.
5. Gilanie, g., et al., *object extraction from t2 weighted brain mr image using histogram based gradient calculation.* Pattern recognition letters, 2013. **34**(12): p. 1356-1363.

6. Ullah, h., et al., *autocorrelation optical coherence tomography for glucose quantification in blood.* Laser physics letters, 2015. **12**(12): p. 125602.
7. Asghar, k., et al., *automatic enhancement of digital images using cubic bã© zier curve and fourier transformation.* Malaysian journal of computer science, 2017. **30**(4): p. 300-310.
8. Janjua, h.u., et al., *classification of liver cirrhosis with statistical analysis of texture parameters.* International journal of optical sciences, 2017. **3**(2): p. 18-25.
9. Bajwa, u.i., et al., *computer-aided detection (cade) system for detection of malignant lung nodules in ct slices-a key for early lung cancer detection.* Current medical imaging, 2018. **14**(3): p. 422-429.
10. Gilanie, g., et al., *classification of normal and abnormal brain mri slices using gabor texture and support vector machines.* Signal, image and video processing, 2018. **12**: p. 479-487.
11. Gilanie, g., et al., *colored representation of brain gray scale mri images to potentially underscore the variability and sensitivity of images.* Current medical imaging reviews, 2018. **14**(4): p. 555-560.
12. Janjua, h.u., a. Jahangir, and g. Gilanie, *classification of chronic kidney diseases with statistical analysis of textural parameters: a data mining technique.* International journal of optical sciences, 2018. **4**(1): p. 1-7.
13. Ullah, h., a. Batool, and g. Gilanie, *classification of brain tumor with statistical analysis of texture parameter using a data mining technique.* International journal of industrial biotechnology and biomaterials, 2018. **4**(2): p. 22-36.
14. Lyons, m., et al. *Coding facial expressions with gabor wavelets*. In *proceedings third ieee international conference on automatic face and gesture recognition*. 1998. Ieee.
15. Gilanie, g., *automated detection and classification of brain tumor from mri images using machine learning methods*. 2019, department of computer science, comsats university islamabad, lahore campus.
16. Gilanie, g., et al., *automated and reliable brain radiology with texture analysis of magnetic resonance imaging and cross datasets validation.* International journal of imaging systems and technology, 2019. **29**(4): p. 531-538.
17. Gilanie, g., et al., *computer aided diagnosis of brain abnormalities using texture analysis of mri images.* International journal of imaging systems and technology, 2019. **29**(3): p. 260-271.
18. Amjad, m., et al., *fourier-transform infrared spectroscopy (ftir) for investigation of human carcinoma and leukaemia.* Lasers in engineering (old city publishing), 2021. **51**.
19. Gilanie, g., et al., *risk-free who grading of astrocytoma using convolutional neural networks from mri images.* Multimedia tools and applications, 2021. **80**(3): p. 4295-4306.
20. Gilanie, g., et al., *coronavirus (covid-19) detection from chest radiology images using convolutional neural networks.* Biomedical signal processing and control, 2021. **66**: p. 102490.
21. Gilanie, g., et al., *riceagenet: age estimation of pakistani grown rice seeds using convolutional neural networks.* International journal of computational intelligence in control, 2021. **13**(2): p. 831-843.
22. Gilanie, g., et al., *ricenet: convolutional neural networks-based model to classify pakistani grown rice seed types.* Multimedia systems, 2021: p. 1-9.
23. Gilanie, g., et al., *digital image processing for ultrasound images: a comprehensive.* Digital image processing, 2021. **15**(3).
24. Malik, s., et al., *enhancing contrast in optical imaging of cancer tissues and study the spectral properties of methylene blue.* Acta microscópica, 2021. **30**(2): p. 49-57.
25. Rafiq, m., et al., *reconstruction of scene using corneal reflection.* Multimedia tools and applications, 2021: p. 1-17.
26. Talaat, f.m., *real-time facial emotion recognition system among children with autism based on deep learning and iot.* Neural computing and applications, 2023. **35**(17): p. 12717-12728.

27. Ghaffar, a.a., et al. *Refined sentiment analysis by ensembling technique of stacking classifier*. In *international conference on soft computing and data mining*. 2022. Springer.
28. Gilanie, g., et al., *an automated and real-time approach of depression detection from facial micro-expressions*. Computers, materials & continua, 2022. **73**(2).
29. Gilanie, g., et al. *Fernet: a convolutional neural networks based robust model to recognize human facial expressions*. In *international conference on soft computing and data mining*. 2022. Springer.
30. Iqbal, m.j., et al., *automatic brain tumor segmentation from magnetic resonance images using superpixel-based approach*. Multimedia tools and applications, 2022. **81**(27): p. 38409-38427.
31. Rubab, s.f., et al. *The comparative performance of machine learning models for covid-19 sentiment analysis*. In *international conference on soft computing and data mining*. 2022. Springer.
32. Ullah, h., et al., *diagnosis of ocular diseases using optical coherence tomography (oct) at λ= 840 nm*. Lasers in engineering (old city publishing), 2022. **53**.
33. Wazir, e., et al. *Early stage detection of cardiac related diseases by using artificial neural network*. In *international conference on soft computing and data mining*. 2022. Springer.
34. Yaseen, m., et al., *in-vitro evaluation of anticancer activity of rhodamine-640 perchlorate on rhabdomyosarcoma cell line*. 2022.
35. Afzal, f., et al., *detection of uric acid in uv-vis wavelength regime*. Journal of nanoscope (jn), 2023. **4**(1): p. 75-81.
36. Ahmed, m., et al., *review of artificial intelligence-based covid-19 detection and a cnn-based model to detect covid-19 from x-rays and ct images*. Vfast transactions on software engineering, 2023. **11**(2): p. 100-112.
37. Asghar, s., et al., *water classification using convolutional neural network*. Ieee access, 2023. **11**: p. 78601-78612.
38. Batool, s.n. and g. Gilanie, *cvip-net: a convolutional neural network-based model for forensic radiology image classification*. Computers, materials & continua, 2023. **74**(1).
39. Ghani, m. And g. Gilanie, *the iomt-based risk-free approach to lung disorders detection from exhaled breath examination*. Intelligent automation and soft computing, 2023. **36**(3): p. 2835-2847.
40. Gilanie, g., et al., *an automated and risk free who grading of glioma from mri images using cnn*. Multimedia tools and applications, 2023. **82**(2): p. 2857-2869.
41. Hafeez, h.a., et al., *a cnn-model to classify low-grade and high-grade glioma from mri images*. Ieee access, 2023. **11**: p. 46283-46296.
42. Khera, e.a., et al., *characterizing nickel oxide thin films for smart window energy conversion applications: combined experimental and theoretical analyses*. Chemistryselect, 2023. **8**(37): p. E202302320.
43. Nazir, a., et al., *exploring breast cancer texture analysis through multilayer neural networks*. Scientific inquiry and review, 2023. **7**(3): p. 32-47.
44. Shafiq, h., et al., *dental radiology: a convolutional neural network-based approach to detect dental disorders from dental images in a real-time environment*. Multimedia systems, 2023. **29**(6): p. 3179-3191.
45. Ullah, h., et al., *proteins and triglycerides measurement in blood under ultraviolet (uv)/visible (vis) spectroscopy at & lambda;= 190 to 1100 nm with an additional he-ne laser source*. Lasers in engineering, 2023. **55**(3-6): p. 157-167.
46. Ullah, h., et al., *assessing graphene oxide (go) and cuo nanocomposites for effective antibacterial properties using laser interferometry*. Lasers in engineering (old city publishing), 2023. **55**.

47. Ullah, h., et al., *early detection of liver, ovary, breast and stomach tumours in the visible (λ= 630 nm) and infrared (ir)(λ= 10.5 to 5.5 μm) wavelength regimes.* Lasers in engineering (old city publishing), 2023. **54**.
48. Abbas, s.n., et al., *comparative evaluation of yolov11 and u-net for automated fruit detection, quality classification, and defect segmentation.* Notulae botanicae horti agrobotanici cluj-napoca, 2026. **54**(3): p. 15337-15337.
49. Afzaal akhtar, m., et al., *deep learning–driven automated grading of astrocytoma from brain mri using an enhanced densenet-169 framework.* Computación y sistemas, 2026. **30**(2): p. 1135.
50. Farooq, a., et al., *conceptual framework for an eeg-driven human brain interface for neural encoding and decoding.* Spectrum of engineering sciences, 2026. **4**(5): p. 2419-2435.
51. Farooq, a., et al., *a color pre-processing method for tumor segmentation using human liver ct images.* Spectrum of engineering sciences, 2026. **4**(6): p. 3159-3179.
52. Farooq, a., et al., *a non-invasive method for brain tumor detection using computer vision and deep learning techniques.* Spectrum of engineering sciences, 2026. **4**(6): p. 2575-2606.
53. Iqbal, m., et al., *automated identification of mango leaf diseases using deep convolutional neural networks.* Polish journal of environmental studies, 2026.
54. Khursheed, a., et al., *a smart helmet to detect anomalies of its users and environment.* Spectrum of engineering sciences, 2026. **4**(6): p. 399-418.
55. Khursheed, a., et al., *deep learning-based cotton pest classification using computer vision.* Spectrum of engineering sciences, 2026. **4**(6): p. 3869-3897.
56. Khursheed, a., et al., *detection of plant leaf diseases using deep learning convolutional neural networks.* Spectrum of engineering sciences, 2026. **4**(3): p. 5591-5615.
57. Majid, s., et al., *an iot-enabled smart helmet for motorcyclist safety: helmet-use enforcement, alcohol interlock and automatic crash alerting.* Spectrum of engineering sciences, 2026. **4**(3): p. 6004-6016.
58. Naseer, s., et al., *soil texture classification from smartphone field imagery under uncontrolled illumination: a benchmark of eight convolutional architectures.* Spectrum of engineering sciences, 2026. **4**(3): p. 6457-6478.
59. Nawaz, a., et al., *a smart irrigation system using the internet of things and machine learning for water-efficient crop management.* Spectrum of engineering sciences, 2026. **4**(3): p. 6017-6029.
60. Saher, a., et al., *parameters optimization of convolutional neural network using particle swarm optimization for mental disorder detection through facial micro-expressions.* Spectrum of engineering sciences, 2026. **4**(3): p. 5935-5953.
61. Sajid, m., et al., *internet of medical things-driven deep learning approach for automated heart sound classification.* Iet biometrics, 2026. **2026**(1): p. 3212328.
62. Shafique, h., et al., *detecting fraudulent labeling of seed samples using computer vision.* Spectrum of engineering sciences, 2026. **4**(3): p. 5564-5590.
63. Shafique, h., et al., *machine vision approach to discriminate and classify normal and disease-affected crop seeds.* Spectrum of engineering sciences, 2026. **4**(3): p. 5511-5534.
64. Yasir, m., et al., *differentiation of metastatic and primary brain tumor using magnetic resonance imaging.* International journal of radiation research, 2026. **24**(1): p. 259-265.
65. Gilanie, g., et al., *a robust method of bipolar mental illness detection from facial micro expressions using machine learning methods.* Intelligent automation & soft computing, 2024. **39**(1).
66. Naveed, s., et al., *drug efficacy recommendation system of glioblastoma (gbm) using deep learning.* Ieee access, 2024.

67. Rashid, m.s., et al., *automated detection and classification of psoriasis types using deep neural networks from dermatology images.* Signal, image and video processing, 2024. **18**(1): p. 163-172.
68. Saher, a., et al., *a deep learning-based automated approach of schizophrenia detection from facial micro-expressions.* Intelligent automation & soft computing, 2024. **39**(6).
69. Ullah, h., et al., *potential application of ceo2/au nanoparticles as contrast agents in optical coherence tomography.* Journal of optoelectronics and advanced materials, 2024. **26**(july-august 2024): p. 307-315.
70. Ullah, h., et al., *measurements of hyperproteinaemia in human blood using laser interferometry: in vitro study.* Lasers in engineering (old city publishing), 2024. **57**.
71. Adnan, m., et al., *ethnicity classification from facial images using deep learning methods.* Spectrum of engineering sciences, 2025. **3**(9): p. 1082-1156.
72. Ahmed, a., et al., *adapting ipv6 and 6lowpan over wifi-based aodv manets for iot applications.* Spectrum of engineering sciences, 2025. **3**(7): p. 1038-1052.
73. Akhtar, m., et al., *harnessing optics and statistics for early detection and prognosis in breast and ovarian cancer.* Lasers in medical science, 2025. **40**(1): p. 279.
74. Amjad, m., et al., *staging of different tumor by utilizing laser guidance (@ 405 nm) in ct scan images along with statistical analysis.* Lasers in engineering, 2025. **59**(4-6): p. 309-325.
75. Amjad, m., et al., *histopathology of malignant tissues using high-resolution microscopy@ 630nm.* Lasers in engineering, 2025. **59**(4-6): p. 295-308.
76. Anwaar, m., et al., *optimizing document classification using modified relative discrimination criterion and rss-elm techniques.* International journal of advanced computer science & applications, 2025. **16**(4).
77. Batool, s.n., et al., *forensic radiology: a robust approach to biological profile estimation from bone image analysis using deep learning.* Biomedical signal processing and control, 2025. **105**: p. 107661.
78. Gilanie, g., et al., *a robust convolutional neural network-based approach for human emotion classification: cross-dataset validation and generalization.* Spectrum of engineering sciences, 2025. **3**(4): p. 782-798.
79. Gilanie, g., et al., *a robust artificial neural network approach for early detection of cardiac diseases.* Spectrum of engineering sciences, 2025. **3**(4): p. 499-511.
80. Gilanie, g., et al., *deep learning-based approach for estimating the age of pakistani-grown rice seeds.* Spectrum of engineering sciences, 2025. **3**(1): p. 557-572.
81. Gilanie, g., et al., *steganographic secret communication using rgb pixel encoding and cryptographic security.* Spectrum of engineering sciences, 2025. **3**(3): p. 323-336.
82. Gilanie, g., et al., *readable text retrieval from noise-influenced documents using image restoration methods.* Spectrum of engineering sciences, 2025. **3**(3): p. 337-360.
83. Gilanie, g., et al., *parameter optimization of autoencoder for image classification using genetic algorithm.* Spectrum of engineering sciences, 2025. **3**(4): p. 201-213.
84. Kashif, a., et al., *forensic radiology: an intelligent method of age and gender estimation from x-ray scanned bones images using convolutional neural networks.* Spectrum of engineering sciences, 2025. **3**(7): p. 440-487.
85. Latif, a., et al., *a vision-free obstacle detection and alert system using smart knee gloves.* Spectrum of engineering sciences, 2025. **3**(6): p. 816-829.
86. Sajid, m., et al., *iomt-enabled noninvasive lungs disease detection and classification using deep learning-based analysis of lungs sounds.* International journal of advanced computer science & applications, 2025. **16**(2).

87. Siddique, m.a., et al., *a multi-modal approach for exploring sarcoma and carcinoma using ftir and polarimetric analysis.* Microscopy research and technique, 2025.
88. Yang, j., et al., *braincnn: automated brain tumor grading from magnetic resonance images using a convolutional neural network-based customized model.* Slas technology, 2025: p. 100334.
89. Howard, a., et al. *Searching for mobilenetv3*. In *proceedings of the ieee/cvf international conference on computer vision*. 2019.
90. Lucey, p., et al. *The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression*. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. 2010. Ieee.
91. Goodfellow, i.j., et al. *Challenges in representation learning: a report on three machine learning contests*. In *international conference on neural information processing*. 2013. Springer.
92. Mollahosseini, a., b. Hasani, and m.h. mahoor, *affectnet: a database for facial expression, valence, and arousal computing in the wild.* Ieee transactions on affective computing, 2017. **10**(1): p. 18-31.
93. Dhall, a., et al. *Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark*. In *2011 ieee international conference on computer vision workshops (iccv workshops)*. 2011. Ieee.
94. Zhang, k., et al., *joint face detection and alignment using multitask cascaded convolutional networks.* Ieee signal processing letters, 2016. **23**(10): p. 1499-1503.