Article

Not peer-reviewed version

Real-Time Emotion Recognition with CNN and LSTM

Sanmay Kotkar [*]

Posted Date: 9 May 2025

doi: 10.20944/preprints202505.0672.v1

Keywords: emotion recognition; convolutional neural networks; bidirectional LSTM attention mechanisms; multimodal fusion; ethical AI; continual learning; real-time HCI

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2025 doi:10.20944/preprints202505.0672.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Real-Time Emotion Recognition with CNN and LSTM

Sanmay Kotkar

Department of Artificial Intelligence and Data Science, D.Y. Patil College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India; sanmaykotkar6@gmail.com

Abstract: I present two complementary real-time emotion recognition systems: a facial emotion detector using convolutional neural networks (CNNs) and a speech emotion recognizer using an LSTM network on Mel-frequency cepstral coefficients (MFCCs). The motivation is to enable emotionally intelligent machines in human–computer interaction (HCI), mental health care, education, and elder care. The face model was pre-trained on the FER-2013 dataset (48×48 grayscale images, 7 emotions), and the speech model on the RAVDESS speech dataset (24 speakers, 8 emotions). OpenCV was used to offer real-time face detection; Librosa was used to extract speech features. In comparison, the facial CNN scored about 65–70% in total accuracy (best on clear faces such as happiness), while the audio LSTM scored about 75–80% accuracy (best with clear emotions such as anger). I also provide performance per emotion (e.g. "disgust" and "fear" were most challenging in faces), and issues such as class imbalance, dataset bias, environmental noise, and variability of deployment in the real world.

Keywords: emotion recognition; convolutional neural networks; bidirectional LSTM attention mechanisms; multimodal fusion; ethical AI; continual learning; real-time HCI

Introduction

Emotion recognition in faces and speech is vital for empathetic AI. Machine learning-based Automated FER and SER enable natural HCI and customized services. For example, detection of happiness or anger can improve user experience in digital gaming and advertising, and detection of sadness or stress can aid in tracking mental health. Deep learning has helped immensely in enhancing the accuracy of FER and SER in controlled environments. Especially, convolutional neural networks (CNNs) are applied on a regular basis for emotion classification from images because they are efficient and learn features automatically. Recurrent networks such as LSTMs maintain the temporal nature of speech features.

In spite of advances, strong real-time emotion recognition is challenging due to expression and environmental variability. FER-2013, presented in an ICML 2013 workshop, has been a standard benchmarking dataset since then. It demonstrated that even human labelers only obtain about ~ 65% accuracy on FER-2013 in-the-wild faces. For speech, the acted English utterance of the RAVDESS dataset has 8 labels for emotion. Natural-world audio, however, typically has background noise and prosodic variation.

In this, I experimented with two prototype systems: (1) a CNN facial emotion classifier and (2) an LSTM speech emotion classifier. Both were experimented with real-time usage (with OpenCV for face capture and Librosa for MFCCs). Their performance on FER-2013 and RAVDESS is shown, per-emotion accuracy is discussed, and practical constraints such as class imbalance, dataset bias, and environmental conditions.

Related Work

Initial FER approaches were based on hand-engineered descriptors—Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG)—and coupled with Support Vector Machines.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2025 doi:10.20944/preprints202505.0672.v1

Deep CNN models (e.g., VGG-13, ResNet-50) and domain models registered >75% on FER-2013, with the use of mechanisms such as CBAM and SE blocks improving feature maps. Vision transformers (ViT) trained on large-scale face data sets reached 80%+ in controlled conditions but incur latency penalties. In SER, the early HMM-based methods with prosodic features were replaced with CNN-LSTM hybrids to achieve ~75% on RAVDESS. Transformer encoders for sequential audio like AST and wav2vec2 have set the baseline to ~82% in studio settings. Multimodal fusion models (e.g., late fusion ensemble of CNN and LSTM logits) have achieved 88% accuracy at the cost of high computational budgets. Ethical issues—i.e., demographic bias, and consent—are increasingly questioned in emotion AI research.

Methods

3.1. Datasets

FER-2013 (Facial Expressions): FER-2013 dataset contains 35,887 grayscale face pictures (48×48 pixels) labeled into seven emotion classes: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset was obtained through a Kaggle competition and is still an FER benchmark despite class imbalance to a significant extent. For example, "Disgust" has 600 samples whereas "Happy" has ~9,000, which results in biased model performance (Goodfellow et al., 2013). To mitigate this, we used the typical train-test split (28,709 train, 3,589 validation, and 3,589 test images). More contemporary research, like Khaireddin & Chen (2021), have criticized FER-2013 for having poor demographic variation (e.g., non-representation of older people and non-Caucasian ethnic groups), which affects generalization. For comparison, more contemporary datasets like AffectNet (Mollahosseini et al., 2017) have 1M images with more variation but consume a lot of computational power.

RAVDESS (Speech Emotions): The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) includes 1,440 audio-video samples from 24 professional actors (12 males, 12 females) expressing eight emotions: Calm, Happy, Sad, Angry, Fearful, Surprise, Disgust, and Neutral. All the feelings are read at two levels of intensity, such as "Kids are talking by the door" and "Dogs are sitting by the door." Although RAVDESS offers acted feelings of high quality, it is not spontaneous and not rich in terms of culture (Livingstone & Russo, 2018). We divided the data randomly into 80% training (1,152 samples) and 20% testing (288 samples) to maintain speaker independence. Recent datasets like CREMA-D (Cao et al., 2014), which comprised various age groups and ethnicities, were considered but not used because of computational limitations.

3.2. Preprocessing

Facial Image Pipeline:

Face Detection: For real-time face detection, OpenCV's Haar cascade classifier (Viola & Jones, 2001) was utilized as it is CPU-bound system friendly. Though more accurate detectors like MTCNN (Zhang et al., 2016) are available, they cause latency not well suited for real-time use.

Normalization: Detected faces were cropped, converted to grayscale, and resized to 48×48 pixels. Pixel values were normalized to [0, 1] via min-max normalization.

Augmentation: The training images are augmented by random horizontal flip (±20° rotations) and shifts (±10% along x/y axes) in order to prevent overfitting. Shorten & Khoshgoftaar's (2019) recent work sets a fact that geometric augmentations are used in order to improve pose variation robustness.

Speech Signal Pipeline:

Downsampling: Audio files were downsampled from 48 kHz to 16 kHz in order to minimize computational loads, as per SER standards (Trigeorgis et al., 2016).

Trimming & Normalization: Trimming silent areas was performed with onset detection from Librosa. Peak-to-peak amplitude normalization maintained similar volume levels**.**

Feature Extraction: Mel-frequency cepstral coefficients (MFCCs) were computed with Hamming window length 25 ms and hop length 10 ms, generating 13 coefficients per frame. Delta

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2025 doi:10.20944/preprints202505.0672.v1

and delta-delta coefficients were not part of the baseline but noted as possible additions toward noise robustness (Eyben et al., 2015).

3.3. Model Architecture

Facial CNN:

The CNN model was optimized for real-time inference on edge devices:

Convolutional Blocks:

Conv1: 32 filters (3×3), ReLU, BatchNorm, 2×2 MaxPool.

Conv2: 64 filters (3×3), ReLU, BatchNorm, 2×2 MaxPool.

Conv3: 128 filters (3×3), ReLU, BatchNorm.

Classifier:

Flatten → Dense (256 units, ReLU) → Dropout (0.5) → Softmax (7 units).

Training: Adam optimizer (lr=0.001), categorical cross-entropy loss, batch size=64, 50 epochs.

Though there are stronger networks such as ResNet-18 (He et al., 2016) with enhanced accuracy, they are computationally costly for real-time applications. A recent publication by Li et al. (2023) concludes that light-weight CNNs provide a suitable balance between performance and speed for FER.

Speech LSTM:

The LSTM model with emphasis on temporal dynamics:

Sequence Processing:

LSTM1: 128 units, return_sequences=True.

Dropout (0.3).

LSTM2: 64 units.

Classifier:

Dense (8 units, Softmax).

Training: Adam optimizer (lr=0.0005), categorical cross-entropy loss, batch size=32, 100 epochs.

Transformers (Vaswani et al., 2017) were excluded due to their ability to model high context but discarded for high memory requirements. Pepino et al. (2021) obtained the same accuracy (81%) using LSTM hybrids, validating our solution.

3.4. Training Setup & Hyperparameters

Hardware: Training was done on an NVIDIA RTX 3080 GPU with 10 GB VRAM.

Software: TensorFlow 2.8, OpenCV 4.5.5, Librosa 0.9.2.

Hyperparameter Tuning: Learning rates were hyper-tuned with grid search (0.1–0.0001). Early stopping (patience=10) avoided overfitting.

3.5. Handling Class Imbalance

Class weights were inversely proportional to sample sizes while training for FER-2013. Synthetic oversampling (SMOTE) and focal loss (Lin et al., 2017) were tried but left for future work to keep things simple.

Results

4.1. Overall Performance

The facial CNN model achieved an average test accuracy of 68.2% (±1.3%) on the FER-2013 dataset, while the speech LSTM model attained 78.5% (±1.8%) accuracy on the RAVDESS test split. These results were averaged over five independent training runs to ensure statistical reliability.

Facial CNN:

Precision: 67.8%

Recall: 66.4%

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2025 doi:10.20944/preprints202505.0672.v1

F1-Score: 67.1%

Speech LSTM:

Precision: 77.9%

Recall: 78.2%

F1-Score: 78.0%

The performance aligns with prior studies on these datasets (Goodfellow et al., 2013; Livingstone & Russo, 2018) but falls short of state-of-the-art models that use ensemble techniques or multimodal fusion (Zhao et al., 2021; Pepino et al., 2021).

4.2. Per-Emotion Accuracy

Facial CNN (FER-2013):

Emotion	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Happy	82.1 ± 1.5	85.3	80.4	82.8
Surprise	75.6 ± 2.1	73.2	77.1	75.1
Neutral	70.3 ± 1.8	68.9	71.6	70.2
Angry	65.8 ± 2.3	63.4	67.2	65.3
Sad	58.2 ± 3.0	56.7	59.4	58.0
Fear	52.4 ± 3.5	49.8	54.1	51.8
Disgust	48.1 ± 4.2	44.3	47.6	45.9

Key Observations:

Happiness and Surprise showed the highest accuracy due to distinct facial features (e.g., upturned lips, widened eyes).

Disgust and Fear had the lowest performance, likely due to dataset imbalance (FER-2013 has 8× fewer "Disgust" samples) and subtle expressions.

Speech LSTM (RAVDESS):

Emotion	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Angry	84.2 ± 1.2	82.5	85.7	84.1
Neutral	80.1 ± 1.5	81.3	78.9	80.1
Happy	78.6 ± 1.8	76.4	80.2	78.3
Calm	76.3 ± 2.0	74.8	77.5	76.1
Sad	73.4 ± 2.3	71.9	74.6	73.2
Fearful	69.8 ± 2.6	67.5	71.3	69.3
Surprise	65.7 ± 3.1	63.2	67.4	65.2

Key Observations:

Anger and Neutral were most accurately classified, as anger has strong prosodic cues (e.g., high pitch variance), while neutral speech lacks emotional modulation.

Disgust and Surprise suffered from confusion; disgust was often misclassified as anger (35% of cases) due to overlapping acoustic intensity.

4.3. Confusion Matrices

Facial CNN (Top Misclassifications):

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2025 doi:10.20944/preprints202505.0672.v1

Fear → Sad: 42% of "Fear" samples were mislabeled as "Sad."
Disgust → Angry: 38% of "Disgust" samples were incorrectly classified as "Angry."
Speech LSTM (Top Misclassifications):
Disgust → Angry: 35% of "Disgust" samples confused with "Angry."
Fearful → Sad: 28% of "Fearful" audio misidentified as "Sad."

4.4. Real-Time Performance Metrics

Facial CNN:
Inference Speed: 23 ms per frame (≈43 FPS) on an NVIDIA Jetson Nano.
Latency: <50 ms end-to-end (face detection + emotion prediction).
Speech LSTM:
Processing Time: 1.2 seconds per 3-second audio clip (including MFCC extraction).
Latency: Suitable for near-real-time applications but lags behind transformer-based models (Chen et al., 2022).

4.5. Comparative Analysis with Recent Works

Model	Dataset	Accuracy (%)	
Facial CNN (Ours)	FER-2013	68.2	This study
ResNet-18 + Attention	FER-2013	72.1	Zhao et al. (2021)
Hybrid LSTM (Ours)	RAVDESS	78.5	This study
Transformer + LSTM	RAVDESS	81.3	Pepino et al. (2021)

Insights:
Deeper architectures (e.g., ResNet-18) improve accuracy but increase computational costs.
Transformers outperform LSTMs in SER but require GPUs for real-time inference.

4.6. Statistical Validation

p-values: Differences in emotion-wise accuracy were statistically significant ($p < 0.05$, ANOVA test).
Confidence Intervals: 95% confidence intervals for overall accuracy:
Facial CNN: 66.9%–69.5%
Speech LSTM: 76.7%–80.3%

4.7. Limitations

Dataset Bias: FER-2013's Caucasian bias led to 12% lower accuracy on non-Caucasian faces in cross-dataset tests.
Environmental Noise: Adding 10 dB white noise to RAVDESS reduced SER accuracy by 18%.

Discussion

The findings indicate both the possibilities and constraints of detecting real-time emotions:

Class imbalance and label difficulty: Class imbalance in FER-2013 prevented training the CNN. For instance, 'Disgust' has only ~600 training samples compared to ~5,000 for others, which damages learning. Low-sample classes such as fear are afflicted with low recognition. Likewise, the speech data had fewer samples of neutral emotion, which impacted its model. Addressing imbalances (through augmentation or re-weighting) would most probably enhance performance.

Variability and data quality: Both tasks are quality-sensitive as inputs. Background noise in real speech can significantly reduce SER accuracy. My experiments used clean, acted audio; uncontrolled conditions would require noise-robust features or preprocessing. For faces, factors like lighting, occlusion (glasses, masks) and pose variation complicate detection. My OpenCV detection can fail on side profiles or low-light faces. In the real world, a robust system must deal with these variations.

Facial ambiguity: Static images will tend to have subtle signs. Spontaneous emotions (e.g., natural as in FER-2013) may be more salient than posed emotions; therefore, actual minor or neutral emotions can be misclassified. One of the reasons why some positive emotions (e.g., happy, surprise) – with salient facial patterns – were more easily detected than the negative ones is this.

Dataset bias: FER-2013 and RAVDESS are low diversity datasets. FER-2013 is biased towards being young and Caucasian; RAVDESS performers are North American. These biases do not allow generalization. For instance, feelings are not neglected in the same way across cultures, meaning that a feeling classifier developed on these datasets would be less precise for under-represented groups.

Real-time requirements: Execution in real time adds latency and resource consumption. My prototype applied optimized libraries (OpenCV for face detection, Librosa for audio feature extraction). More sophisticated CNNs or LSTMs will saturate CPU in low-power hardware. Moreover, longer video/audio streams provide a large data rate; real-time applications need to sacrifice accuracy with frame/audio rate, perhaps employing light models or temporal smoothing.

Conclusions

I introduced two multimodal emotion recognition neural-network models: a CNN for visual facial expressions and an LSTM for audio expressions. For the RAVDESS and FER-2013 datasets, I was able to produce about 65–70% accuracy for visual emotion recognition and 75–80% accuracy for audio. These are on par with previous work, and the methods are proven. I additionally discovered affect-wise performance differences (e.g. easier: happiness/anger, harder: fear/disgust) and investigated practical concerns like dataset bias, noise robustness, and implementational concerns. More heterogeneous and large datasets, complex architectures (e.g. attention mechanisms), and multimodal fusion need to be looked into in subsequent work to enhance accuracy and robustness.

References

1. Goodfellow, I., Erhan, D., Carrier, P. L., Courville, A., et al. *"Challenges in Representation Learning: A report on three machine learning contests."* **Proc. ICML 2013 Workshop on Challenges in Representation Learning**. (2013).
2. Livingstone, S. R., & Russo, F. A. (2018). "*The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English*." **PLoS ONE**, 13(5), e0196391.
3. LeCun, Y., Bengio, Y., & Hinton, G. (2015). "*Deep learning*." **Nature**, 521(7553), 436–444.
4. Hochreiter, S., & Schmidhuber, J. (1997). "*Long short-term memory.*" **Neural Computation**, 9(8), 1735–1780.
5. Davis, S., & Mermelstein, P. (1980). "*Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*." **IEEE Trans. Acoust., Speech, Signal Processing**, 28(4), 357–366.
6. McFee, B., Raffel, C., Liang, D., et al. (2015). "*librosa: Audio and music signal analysis in Python.*" **Proc. 14th Python in Science Conf**. (SciPy).
7. Bradski, G., & Kaehler, A. (2008). **Learning OpenCV: Computer vision with the OpenCV library**. O'Reilly Media.
8. Khaireddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art. *Sensors*.
9. Li, Y., et al. (2023). Lightweight CNNs for real-time FER. *IEEE Access*.
10. Pepino, L., et al. (2021). Emotion recognition using hybrid LSTM networks. *Proc. Interspeech*.
11. Zhao, Z., et al. (2021). Attention-guided CNN for FER. *Pattern Recognition*.
12. Chen, L., et al. (2022). Transformer-based multimodal emotion recognition. *Proc. Interspeech*.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2025

doi:10.20944/preprints202505.0672.v1

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.