International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Deep Learning Framework for Speech Emotion Recognition: A Gender-Aware Hierarchical Pipeline with Optimized 18-Layer Convolutional Neural Network

Dr. Savita Jain

Assistant Professor, Department of Computer Science, Apex University, Jaipur, Rajasthan.

Abstract

The field of Affective Computing has emerged as a crucial domain in human-computer interaction, with Speech Emotion Recognition (SER) serving as a cornerstone for developing intuitive, context-aware systems. While traditional Automated Speech Recognition (ASR) frameworks have achieved considerable maturity in decoding semantic content, recognizing the underlying emotional state from spoken language remains a computationally complex challenge. Real-world acoustic signals are heavily influenced by environmental noise, speaker idiosyncrasies, and physical variability across genders. This paper introduces a high-performance, structurally optimized hierarchical framework that addresses these limitations through three primary contributions: (1) a dense 182-feature extraction pipeline unifying spectral, linear predictive, dynamic energy, prosodic, and statistical shape profiles; (2) an early-stage, gender-aware hierarchical pipeline driven by a Gender Recognition (GR) circuit that splits the processing stream based on fundamental frequency distribution to eliminate cross-gender acoustic overlaps; and (3) a customized 18-layer Deep Convolutional Neural Network (CNN) integrated with meta-heuristic hyper-parameter optimization. The system is evaluated on the RAVDESS and SAVEE benchmark corpora, demonstrating superior multi-class emotion classification accuracy and operational efficiency compared to baseline Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) architectures.

Keywords: Speech Emotion Recognition, Convolutional Neural Network, Gender Recognition Circuit, Affective Computing, RAVDESS, SAVEE, Feature Extraction, Deep Learning.

1. Introduction

The field of Affective Computing has grown rapidly over the past two decades, driven by the widespread deployment of intelligent voice assistants, telehealth platforms, and human-robot interaction systems. Speech Emotion Recognition (SER) constitutes a foundational pillar of these technologies, seeking to decode the emotional state of a speaker from raw audio signals rather than from the semantic content of words. While Automatic Speech Recognition (ASR) systems now achieve near-human accuracy on clean speech benchmarks, the reliable identification of emotional subtext—the "how" rather than the "what" of spoken language—remains a highly complex, unresolved computational challenge [1, 28].

Real-world acoustic signals pose significant processing difficulties. Environmental noise, varying recording conditions, and substantial physiological differences between male and female vocal tracts

introduce high variability into raw waveform inputs. Conventional frameworks addressing this challenge have historically relied on shallow machine learning architectures such as Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), and Support Vector Machines (SVMs). While these methods established foundational baselines, they suffer from heavy dependence on manual feature engineering and steep performance degradation when deployed in conditions outside highly controlled laboratory environments [2].

The emergence of deep learning introduced architectures capable of automatically learning hierarchical feature representations from raw data [16]. Multi-Layer Perceptrons (MLPs) and Recurrent Neural Networks (RNNs), refined into Long Short-Term Memory (LSTM) blocks, improved the modelling of temporal dependencies in speech. However, these architectures encounter clear performance bottlenecks when processing high-dimensional, hybrid feature matrices, exhibiting high computational latency and an inability to capture multi-dimensional spatial patterns within acoustic spectrograms [3].

A critical and underaddressed research deficit persists in current literature: the overwhelming majority of public benchmarks and state-of-the-art SER architectures are built and optimized exclusively for Western languages such as English and German. Standardized, high-fidelity emotional speech databases tailored to regional Indian languages such as Hindi and Telugu remain severely scarce. South Asian linguistic profiles rely on distinct syllabic pacing, unique tonal accents, and localized dialectal variations—parameters for which Western-trained deep models exhibit significant performance drops upon deployment [4].

This paper addresses these limitations through three primary contributions:

1. A dense 182-feature extraction pipeline that unifies spectral, linear predictive, dynamic energy, prosodic, and statistical shape profiles to capture micro-level emotional nuances.
2. An early-stage, gender-aware hierarchical pipeline driven by a Gender Recognition (GR) circuit that splits the processing stream based on fundamental frequency distribution to eliminate cross-gender acoustic overlaps.
3. A customized 18-layer Deep Convolutional Neural Network (CNN) [22] integrated with meta-heuristic hyper-parameter optimization to boost classification stability and multi-class accuracy across global benchmark datasets.

2. Literature Review and Theoretical Framework

2.1 Evolution of Affective Computing and Acoustic Modelling

The quest to automate Speech Emotion Recognition has evolved through distinct technological eras over the past three decades. Early pioneering frameworks in affective computing relied heavily on handcrafted feature engineering paired with traditional rule-based statistical classifiers. Models such as HMMs, GMMs, and SVMs served as industry standards for evaluating vocal architecture. Scherer (2003) proposed foundational acoustic correlates of emotion, establishing the groundwork for feature-centric approaches [5]. While these traditional frameworks provided a foundational baseline, they exhibited significant limitations when shifted outside highly controlled laboratory environments, struggling to generalize across diverse speaker populations and demonstrating steep accuracy declines in cross-corpus testing scenarios [6].

2.2 Deep Learning Architectures in Speech Analytics

The emergence of Deep Learning shifted the paradigm of speech processing by introducing architectures capable of automatically learning complex, hierarchical feature representations directly from raw data.

CNN architectures applied directly to spectrogram representations have demonstrated strong discriminative capacity for emotion categories [24]. Trigeorgis et al. (2016) subsequently demonstrated that end-to-end CNN models could outperform hand-engineered feature methods on SER benchmarks [7]. Despite their theoretical suitability, basic LSTMs and shallow MLPs hit clear performance bottlenecks when processing high-dimensional hybrid feature matrices. These architectures are prone to high computational latency, require intensive memory allocations, and frequently fail to capture multi-dimensional, localized spatial patterns found within acoustic spectrograms. This paved the way for CNNs, which treat speech representations similarly to visual feature maps, maximising pattern recognition while reducing operational latency [8, 18].

2.3 Critical Analysis of the Regional Linguistic Deficit

While modern deep architectures have drastically advanced multi-class emotion recognition on international benchmarks, a systemic gap remains in global affective computing literature. The overwhelming majority of high-fidelity, publicly accessible emotional speech datasets—such as RAVDESS, SAVEE, and Emo-DB [19]—are designed and recorded exclusively using Western languages. Consequently, state-of-the-art deep learning models trained on these corpora inherently inherit the phonetic, syntactic, and structural constraints of Western speech mechanics. Regional Indian languages such as Hindi and Telugu are built upon fundamentally different acoustic parameters, characterised by unique syllabic pacing, intricate tonal inflections, and distinct dialectal variations. When deep networks optimised for Western speech profiles are deployed in native Indian contexts, they suffer significant performance degradation due to these unmapped acoustic variances [9, 10]. This research provides a crucial step toward architectures with the structural adaptability required to handle diverse acoustic spaces.

2.4 Summary of Key Prior Works

Table 1: Summary of Key Prior Works in Speech Emotion Recognition

Author(s) & Year	Method Used	Dataset	Key Finding / Limitation
Scherer (2003)	Acoustic Feature Analysis (HMM)	Controlled Lab Corpus	Established prosodic correlates of emotion; limited to lab conditions
Trigeorgis et al. (2016)	End-to-End CNN	AVEC 2016	Outperformed handcrafted features; lacked gender-awareness
Zhao et al. (2019)	LSTM + CNN Hybrid	IEMOCAP [20]	Improved temporal modelling; high computational overhead
Mirsamadi et al. (2017) [17] further demonstrated that locally attended RNN models substantially improve temporal emotion	Attention-based CNN	RAVDESS, SAVEE	High accuracy on English corpora; no regional language coverage

Author(s) & Year	Method Used	Dataset	Key Finding / Limitation
modelling. Kwon et al. (2021)			
Proposed Framework	18-Layer CNN + GR Circuit + PCA	RAVDESS, SAVEE	Gender-aware pipeline; superior accuracy and latency

3. Proposed Methodology

A high-performance, robust SER framework requires a careful balance between high-fidelity feature extraction and optimised neural architecture. The proposed system addresses the limitations of traditional shallow networks by implementing a dual-stage, hierarchical pipeline. First, a Gender Recognition (GR) circuit splits incoming speech signals based on pitch characteristics. Subsequently, a customised, deep 18-layer CNN handles multi-class emotion classification. The complete system pipeline is summarised in Figure 1 below.

Figure 1: System Architecture Pipeline of the Proposed SER Framework

Stage	Component	Description
1	Raw Audio Input	Waveform (.wav) input sampled at 22,050 Hz
2	Pre-processing	Pre-emphasis filtering → Framing & Windowing → Syllable Segmentation
3	Feature Extraction	182-feature dense vector (spectral, prosodic, linear predictive, statistical)
4	Gender Recognition Circuit	Pitch PDF mean → Male / Female decision fork
5a	Male CNN Stream	Targeted 18-layer CNN optimised for male vocal frequency range
5b	Female CNN Stream	Targeted 18-layer CNN optimised for female vocal frequency range
6	Softmax Output Layer	8-class emotion label: Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised, Neutral

3.1 Signal Pre-processing and Syllable Segmentation

To minimise environmental noise, speaker variability, and external recording factors, raw audio wave files undergo a rigorous three-stage pre-processing pipeline before features are extracted.

3.1.1 Pre-emphasis Filtering

The acoustic signal is passed through a first-order FIR high-pass filter to flatten the signal spectrum and balance the energy of high-frequency formants. The discrete-time difference equation governing this filter

is expressed as:

$$y[n] = x[n] - \alpha \times x[n-1], \quad 0.95 \leq \alpha \leq 0.97 \qquad (1)$$

Where x[n] is the input speech sample and α is the pre-emphasis coefficient, typically set to 0.97 to attenuate low-frequency dominance and enhance high-frequency phonetic energy.

3.1.2 Framing and Windowing

Because speech is a non-stationary signal, the pre-emphasised signal is divided into short, quasi-stationary frames of between 20 ms and 30 ms with a 50% overlap. To eliminate artificial discontinuities at frame edges, each frame is multiplied by a Hamming window function:

$$w[n] = 0.54 - 0.46 \times \cos(2\pi n \div (N-1)), \quad 0 \leq n \leq N-1 \qquad (2)$$

Where N represents the total number of discrete samples inside each isolated frame. The Hamming window function smoothly tapers the signal at both ends of each frame, significantly reducing spectral leakage artefacts during subsequent Fourier Transform analysis.

3.1.3 Automated Syllable-like Segmentation

Rather than computing statistics indiscriminately across silent gaps and unvoiced intervals, an automated syllable-like segmentation framework is implemented. The algorithm detects intense energy bursts and vocalic nuclei by applying a dynamic energy threshold. Feature calculations are thus focused exclusively on active, emotionally expressive speech segments, reducing the noise-to-signal ratio in the final feature matrix.

3.2 Dense 182-Feature Extraction Profile

To capture subtle emotional cues at multiple acoustic levels, the system extracts a dense vector of 182 distinct features combining spectral, prosodic, linear predictive, and statistical metrics across every active frame. The full breakdown is presented in Table 2 below.

Table 2: Dense 182-Feature Extraction Profile for Emotion Classification

Feature Category	Count	Specific Acoustic Metrics Included	Mathematical / Physical Relevance
Spectral Vectors	39	Mel Frequency Cepstral Coefficients (MFCCs 1–13) [25], Delta (Velocity), Delta-Delta (Acceleration)	Emulates non-linear human hearing via log-compression and Inverse Discrete Cosine Transform (IDCT) [11]
Linear Predictive Profiles	26	Linear Prediction Cepstrum Coefficients (LPCC)	Models the physical resonance and shape of the human vocal tract filter
Dynamic Energy Sub-bands	24	Mel Energy Spectrum Dynamic Coefficients (MEDC)	Measures energy variations across log-spaced frequency bands over time
Prosodic Elements	45	Pitch / Fundamental Frequency (f_0), Intensity, Duration trajectories, Tempo markers	Tracks global intonation changes, excitement levels, and speech stress patterns

Feature Category	Count	Specific Acoustic Metrics Included	Mathematical / Physical Relevance
Statistical & Shape Dimensions	48	Kurtosis, Skewness, Formant Dispersion Ranges, Bark scale sub-band energy variance	Captures structural irregularities and morphological shifts in the audio waveform
Total	**182**	—	Complete acoustic profile for robust multi-class emotion classification

Processing a 182-dimensional vector can cause high computational latency and overfitting, a phenomenon known as the "curse of dimensionality". To resolve this, the framework applies Principal Component Analysis (PCA) [13] alongside mutual information metric rankings. This procedure filters out redundant and noisy data points while retaining only highly discriminative feature combinations, significantly reducing system memory requirements and processing overhead.

3.3 Hierarchical System Pipeline and Gender Recognition (GR) Circuit

A critical challenge in Speech Emotion Recognition is cross-gender acoustic overlap: a neutral male voice can exhibit spectral characteristics structurally similar to a sad female voice in a unified feature space. To resolve this, the paper introduces a Gender-Aware Hierarchical Pipeline.

The Gender Recognition (GR) Circuit functions as follows. The preprocessed acoustic vectors first enter a specialized classification node that extracts the Pitch Probability Density Function (PDF) mean. Because male and female fundamental frequencies (F_0) naturally cluster in separate, distinct ranges—typically 85–180 Hz for males and 165–255 Hz for females—this circuit achieves highly accurate early gender sorting. Once sorted, the feature vector is routed to a dedicated, gender-tuned emotion classifier stream, significantly boosting overall system accuracy and reducing processing time by eliminating inter-gender spectral confusion.

Table 3: Gender Recognition Circuit — Fundamental Frequency Threshold Parameters

Parameter	Male Range	Female Range
Fundamental Frequency (F_0)	85 – 180 Hz	165 – 255 Hz
Average Formant F1	270 – 730 Hz	310 – 860 Hz
Average Formant F2	840 – 2290 Hz	920 – 2760 Hz
Pitch PDF Mean Threshold	≤ 165 Hz → Male stream	> 165 Hz → Female stream

3.4 Optimized 18-Layer Deep CNN Architecture

The proposed architecture uses an optimised, high-fidelity 18-layer Deep Convolutional Neural Network (CNN) that outperforms standard baseline networks. The exact sequence of layers, structural dimensions, and regularisers is detailed in Table 4 below.

Table 4: Layer-by-Layer Architecture of the Proposed 18-Layer Deep CNN

Layer No.	Layer Type	Configuration / Parameters	Function
1–3	Conv2D + Batch Normalisation [14] + ReLU	Kernel: 3×3, Filters: 32	Low-level micro-structural acoustic feature map extraction
4	Max Pooling	Stride: 2×2	Spatial downsampling; retains critical activation points
5–10	Deep Conv2D Blocks + Dropout	Filters: 64→256; Dropout [15]: p = 0.30	Deep pattern matching; dropout prevents overfitting
11	Global Average Pooling	1D vector flattening	Converts 2D feature maps into unified 1D feature vector
12–16	Dense Fully-Connected Layers	Units: 512→256→128→64→32	Hierarchical reasoning and feature abstraction
17	Meta-Heuristic Hyper-parameter Optimisation	Dynamic weight / learning-rate tuning	Avoids local minima; optimises network convergence
18	Softmax Output Classification	8 nodes (one per emotion class)	Probability distribution across 8 discrete emotional states

The eight target emotional states classified by the Softmax output layer are: Calm, Fearful, Sad, Surprised, Angry, Disgust, Happy, and Neutral. Meta-heuristic hyper-parameter optimisation at Layer 17 integrates evolutionary search strategies to dynamically adjust weights, learning rates, and internal biases during training, effectively preventing the network from converging to suboptimal local minima.

4. Experimental Results and Discussion

4.1 Experimental Environment and Benchmarking Corpora

The performance, reliability, and generalisation capabilities of the proposed 18-layer deep CNN architecture were rigorously validated using two globally recognised, standard emotional speech corpora.

4.1.1 RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a highly controlled, professional corpus consisting of speech files recorded by 24 professional actors (12 male, 12 female). The database covers 8 distinct emotional states: Calm, Fearful, Sad, Surprised, Angry, Disgust, Happy, and Neutral. A total of 1,440 audio-only speech files were used in this study, with a 70:15:15 train/validation/test split.

4.1.2 SAVEE Dataset

The Surrey Audio-Visual Expressed Emotion (SAVEE) database features recordings from 4 native English male actors [12]. It provides a reliable benchmark for evaluating model performance on highly precise vocal expressions across multi-class emotion profiles. A total of 480 utterances were used for evaluation, with the same 70:15:15 split strategy.

4.1.3 Data Augmentation Procedures

To simulate real-world, naturalistic conditions and prevent overfitting, raw audio samples were subjected to systematic data augmentation. This included: (a) introduction of synthetic Gaussian white noise at signal-to-noise ratios of 10 dB, 20 dB, and 30 dB [26]; (b) variable speed-stretch factors of ±10% and ±20%; and (c) pitch-shifting variations of ±2 semitones. Standard data augmentation techniques including speed perturbation and pitch shifting have proven effective across SER benchmarks [21]. The augmentation strategy effectively tripled the training corpus size and ensured network resilience to common real-world acoustic distortions.

4.2 Performance Metrics

The classification performance was assessed using four standard metrics: Accuracy (A), Weighted Precision (P), Weighted Recall (R), and Weighted F1-Score (F1). The macro-averaged Unweighted Average Recall (UAR) was also computed to account for class imbalance. The formulae used are:

$$\text{Accuracy (A)} = (TP + TN) \div (TP + TN + FP + FN) \quad (3)$$

$$\text{Precision (P)} = TP \div (TP + FP) \quad (4)$$

$$\text{Recall (R)} = TP \div (TP + FN) \quad (5)$$

$$\text{F1-Score} = 2 \times (P \times R) \div (P + R) \quad (6)$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives for each respective emotion class.

4.3 Comparative Benchmarking Against Baseline Models

To confirm the advantages of the proposed 18-layer CNN framework, it was benchmarked against traditional baseline deep learning models under identical training, validation, and testing splits. The comparative performance data is presented in Table 5.

Table 5: Multi-Class Emotion Classification Performance Comparison

Architecture	Feature Space	Corpus	Accuracy (%)	F1-Score (%)	Latency
Multi-Layer Perceptron (MLP)	182-Feature Raw Vector	RAVDESS	71.4	69.8	Medium
Multi-Layer Perceptron (MLP)	182-Feature Raw Vector	SAVEE	68.7	66.9	Medium
Long Short-Term Memory (LSTM)	182-Feature Raw Vector	RAVDESS	76.2	74.5	High
Long Short-Term Memory (LSTM)	182-Feature Raw Vector	SAVEE	73.1	71.3	High
Proposed 18-Layer CNN (Ours)	**PCA-Optimised Stream**	**RAVDESS**	**91.3**	**90.7**	**Low**
Proposed 18-Layer CNN (Ours)	**PCA-Optimised Stream**	**SAVEE**	**88.6**	**87.9**	**Low**

The results clearly demonstrate that the baseline MLP and LSTM models struggled to parse the massive 182-feature dimensionality, leading to higher processing latency and comparatively lower accuracy scores. In contrast, the proposed architecture uses PCA dimensionality reduction to isolate highly discriminative features. By routing these optimised vectors through gender-targeted streams, the system minimises computational overhead while achieving superior classification accuracy across both benchmark datasets. A relative accuracy improvement of 19.9 percentage points over MLP and 15.1 percentage points over LSTM is achieved on the RAVDESS corpus.

4.4 Per-Class Emotion Recognition Performance (RAVDESS)

The per-class performance breakdown on the RAVDESS dataset reveals consistent recognition across all eight emotion classes, as reported in Table 6. The Gender Recognition Circuit demonstrably reduces cross-class confusion between structurally similar emotion profiles.

Table 6: Per-Class Emotion Recognition Performance on the RAVDESS Corpus

Emotion Class	Precision (%)	Recall (%)	F1-Score (%)
Calm	93.1	92.4	92.7
Happy	89.5	91.2	90.3
Sad	90.8	88.7	89.7
Angry	92.3	93.5	92.9
Fearful	88.2	87.6	87.9
Disgust	91.7	90.1	90.9
Surprised	89.0	90.8	89.9
Neutral	93.4	92.8	93.1
Weighted Average	**91.0**	**90.9**	**90.7**

5. Conclusion, Social Significance, and Future Scope

5.1 Conclusion

This paper has presented a structurally optimised, high-performance hierarchical framework for Speech Emotion Recognition that successfully bridges the gap between complex multi-class feature spaces and deep neural processing architectures. By implementing an automated syllable-like segmentation routine and a dense 182-feature extraction pipeline encompassing spectral, prosodic, linear predictive, and statistical shape attributes, the system captures subtle emotional cues with high fidelity. The integration of an early-stage Gender Recognition (GR) circuit effectively mitigates the long-standing challenge of cross-gender acoustic overlap [27], routing optimised feature matrices through dedicated, gender-tuned channels.

When evaluated against the globally recognised RAVDESS and SAVEE corpora, the proposed 18-layer Deep CNN—enhanced by meta-heuristic hyper-parameter optimisation—demonstrated superior convergence stability and outperformed traditional baseline MLP and LSTM architectures in both multi-class classification accuracy and operational latency. Overall accuracies of 91.3% and 88.6% were

achieved on RAVDESS and SAVEE respectively, representing the best results reported under equivalent experimental conditions.

5.2 Social Significance and Real-World Applications

The practical deployment horizons for a robust, minimised-latency SER framework extend across several vital domains of modern society:

- Healthcare Diagnostics and Mental Health Screening: The architecture can be integrated into remote telehealth platforms to assist clinicians in screening for clinical depression, severe anxiety, and postpartum psychological changes by evaluating micro-prosodic variations in patient speech over time.
- Assistive Technologies for Neurodivergence: The system can serve as a core computational engine for real-time emotional assistive applications designed for individuals with Autism Spectrum Disorder (ASD), facilitating clearer interpretation of conversational emotional dynamics.
- Automotive Safety and Driver Monitoring: By embedding this optimised model into in-vehicle communication units, automotive safety suites can actively monitor driver stress levels, voice-based fatigue markers, and sudden road-rage responses to issue preventative safety alerts.
- Intelligent E-Tutoring Environments: In digital learning ecosystems, the framework can analyse student vocal feedback to gauge engagement, confusion, or frustration, enabling adaptive instructional delivery that dynamically adjusts content pacing based on detected affective states.

5.3 Future Scope

While the proposed 18-layer deep CNN architecture delivers outstanding classification metrics on established English-language databases, future research trajectories must aggressively target the foundational research deficit concerning native South Asian linguistic environments. Transfer learning approaches [23] from large multilingual pre-trained models present a promising pathway for cross-lingual SER adaptation. The immediate next phase of this research will focus on the systematic curation, annotation, and structural standardisation of high-quality, publicly accessible emotional speech corpora tailored specifically to regional Indian languages, including Hindi and Telugu. By developing localised acoustic models that inherently encode regional syllable pacing, unique tonal shifts, and distinct dialectal variations, the affective computing landscape can advance toward universally deployable, cross-lingual SER systems capable of serving diverse global populations.

References

1. Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., Taylor J.G., "Emotion Recognition in Human-Computer Interaction", IEEE Signal Processing Magazine, 2001, 18 (1), 32–80.
2. Schuller B., Steidl S., Batliner A., "The INTERSPEECH 2009 Emotion Challenge", Proceedings of INTERSPEECH, 2009, 312–315.
3. Trigeorgis G., Ringeval F., Brueckner R., Marchi E., Nicolaou M.A., Schuller B., Zafeiriou S., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network", Proceedings of ICASSP, 2016, 5200–5204.
4. Koolagudi S.G., Rao K.S., "Emotion Recognition from Speech: A Review", International Journal of Speech Technology, 2012, 15 (2), 99–117.
5. Scherer K.R., "Vocal Communication of Emotion: A Review of Research Paradigms", Speech Communication, 2003, 40 (1–2), 227–256.

6. Kwon S., "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition", Sensors, 2020, 20 (1), 183.
7. Zhao J., Mao X., Chen L., "Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks", Biomedical Signal Processing and Control, 2019, 47, 312–323.
8. Hochreiter S., Schmidhuber J., "Long Short-Term Memory", Neural Computation, 1997, 9 (8), 1735–1780.
9. Srinivasan V., Chakraborty P., "SER Challenges in Indian Language Contexts: A Systematic Review", Journal of Intelligent Systems, 2022, 31 (1), 445–462.
10. Livingstone S.R., Russo F.A., "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions", PLOS ONE, 2018, 13 (5), e0196391.
11. Davis S.B., Mermelstein P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28 (4), 357–366.
12. Jackson P., Haq S., "Surrey Audio-Visual Expressed Emotion (SAVEE) Database", University of Surrey, 2014. https://kahlan.eps.surrey.ac.uk/savee/
13. Pearson K., "On Lines and Planes of Closest Fit to Systems of Points in Space", Philosophical Magazine, 1901, 2 (11), 559–572.
14. Ioffe S., Szegedy C., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", Proceedings of the International Conference on Machine Learning (ICML), 2015, 448–456.
15. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research, 2014, 15 (1), 1929–1958.
16. LeCun Y., Bengio Y., Hinton G., "Deep Learning", Nature, 2015, 521 (7553), 436–444.
17. Mirsamadi S., Barsoum E., Zhang C., "Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention", Proceedings of ICASSP, 2017, 2227–2231.
18. Fayek H.M., Lech M., Cavedon L., "Evaluating Deep Learning Architectures for Speech Emotion Recognition", Neural Networks, 2017, 92, 60–68.
19. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B., "A Database of German Emotional Speech", Proceedings of INTERSPEECH, 2005, 1517–1520.
20. Busso C., Bulut M., Lee C.C., Kazemzadeh A., Mower E., Kim S., Chang J.N., Lee S., Narayanan S.S., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database", Language Resources and Evaluation, 2008, 42 (4), 335–359.
21. Park D.S., Chan W., Zhang Y., Chiu C.C., Zoph B., Cubuk E.D., Le Q.V., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition", Proceedings of INTERSPEECH, 2019, 2613–2617.
22. He K., Zhang X., Ren S., Sun J., "Deep Residual Learning for Image Recognition", Proceedings of IEEE CVPR, 2016, 770–778.
23. Pan S.J., Yang Q., "A Survey on Transfer Learning", IEEE Transactions on Knowledge and Data Engineering, 2010, 22 (10), 1345–1359.
24. Huang Z., Dong M., Mao Q., Zhan Y., "Speech Emotion Recognition Using CNN", Proceedings of ACM Multimedia, 2014, 801–804.

25. Ancilin J., Milton A., "Improved Speech Emotion Recognition with Mel Frequency Magnitude Coefficient", Applied Acoustics, 2021, 179, 108046.
26. Alam M.J., Kenny P., Bhattacharya G., Stafylakis T., "Development of Realistic Emotional Speech Corpus by Mixing Clean Speech with Noisy Speech", Proceedings of INTERSPEECH, 2013, 136–140.
27. Latif S., Rana R., Khalifa S., Jurdak R., Epps J., Schuller B.W., "Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition", IEEE Transactions on Affective Computing, 2021, 12 (4), 1003–1016.
28. Abbaschian B.J., Sierra-Sosa D., Elmaghraby A., "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models", Sensors, 2021, 21 (4), 1249.