Speech Emotion Recognition with Combination of CNN-LSTM and DNN

Internship Training(EC 402)

Name: PONNA DINESH

Admission Number: U20EC009

Venue: Electronics and Communication Department

Duration: 01/01/2024 to 26/04/2024

Guided By:

Dr Suman Deb
(Assistant Professor, DoECE)

PRESENTATION OUTLINE


INTRODUCTION


OBJECTIVE


LITERATURE REVIEW


METHODOLOGY


RESULTS & DISCUSSION


CONCLUSION


REFERENCES

INTRODUCTION

- Speech Emotion Recognition (SER) is a field within natural language processing that aims to detect emotions from spoken language.
- It plays a crucial role in various applications such as human-computer interaction and sentiment analysis.

Applications of SER

- Human-Computer Interaction
- Health care
- Education
- Customer services
- Market Research
- Security

OBJECTIVE

- Feature Extraction
- Model Development
- Model Evaluation
- Comparative Analysis
- Emotional Classifications
- Optimization

LITERATURE REVIEW

- Mishra et al. [1] advances speech emotion recognition by utilizing a combination of MFCC, spectrograms, and mel-spectrograms as inputs to CNNs and DNNs. This innovative fusion approach yielded significant performance enhancements across multiple datasets.
- Bertero et al. [2] enhance interactive dialogue systems by integrating CNN models that detect user emotion and sentiment in real-time, bypassing traditional feature engineering. Their method surpasses feature-based SVM classification with an improved accuracy of 65.7% across six emotion categories.
- Yadav et al. [3] introduce a language-independent framework for emotion classification using a novel combination of CNN and Bi-LSTM. This approach effectively captures both MFCC-based and high-level features, enhancing emotional context analysis. The system's efficacy was validated against multilingual datasets, EmoDB for German and RAVDESS for English.

[1] S. Mishra, P. Warule, and S. Deb, “Speech emotion classification using feature level and classifier-level fusion,” Evolving Systems, vol. 15, pp. 1–14, 11 2023
[2] D. Bertero, F. Siddique, C.-S. Wu, Y. Wan, R. Chan, and P. Fung, “Real-time speech emotion and sentiment recognition for interactive dialogue systems,” 012016, pp. 1042–1047
[3] A. Yadav and D. K. Vishwakarma, “A multilingual framework of cnn and bi-lstm for emotion classification,” in 2020 11th international conference on computing, communication and networking technologies (ICCCNT). IEEE, 2020, pp. 1–6

Methodology and Algorithm

- Input speech signal (EmoDB)

- MFCC
- Delta and Delta-Delta MFCC
- Mel-Spectrogram
- Spectrogram
- Chroma
- Tonnetz

- CNN+LSTM
- DNN

- Emotion recognised

EmoDB Dataset

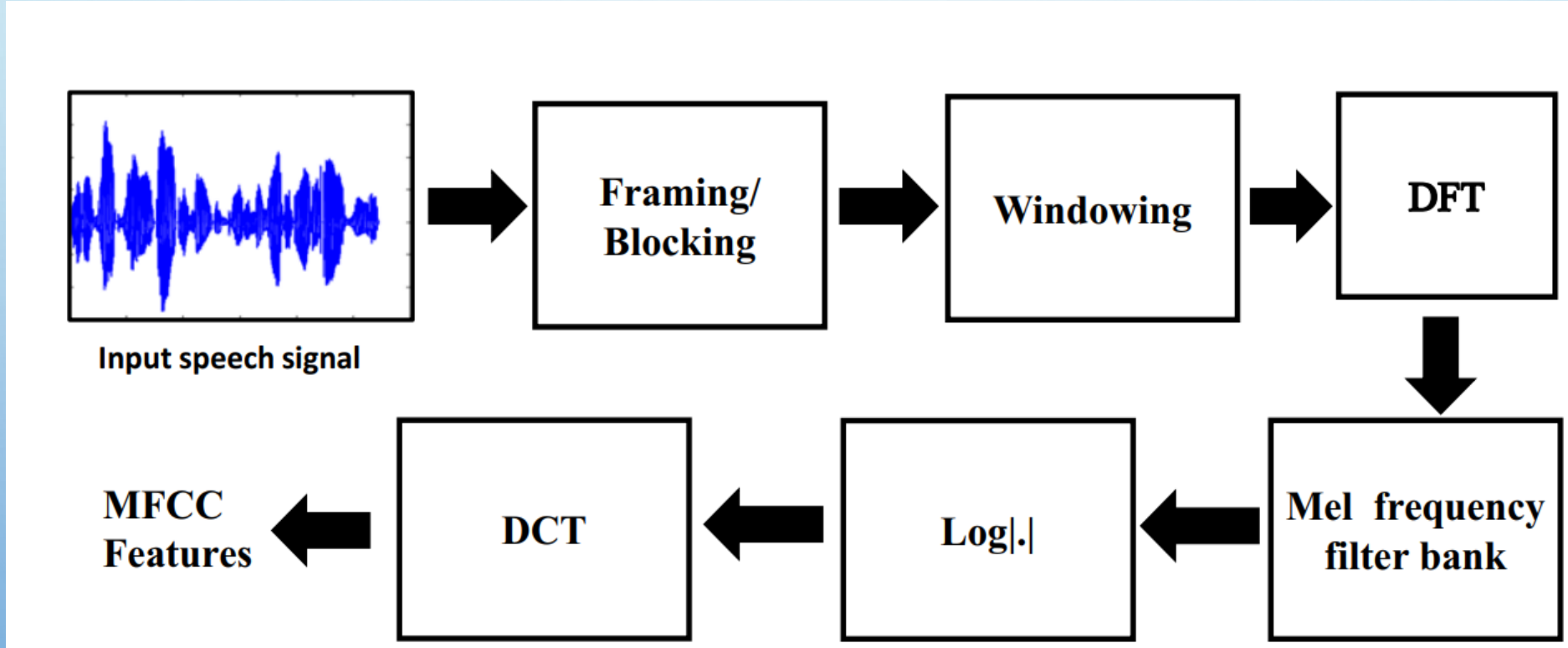
- Anger
- Boredom
- Disgusted
- Anxiety
- Happiness
- Sadness
- Neutral

Speech Pre-processing

- Preprocessing Includes:
  - Pre-emphasis
  - Normalization
  - Framing/Segmentation

Feature Extraction

MFCC



Source: S. Mishra, P. Warule, and S. Deb, "Speech emotion classification using feature level and classifier-level fusion," Evolving Systems, vol. 15, pp. 1–14, 11 2023

Mel Spectrogram

- A mel spectrogram is a visual representation of the frequency spectrum of sounds, scaled to the human ear's perception of pitch, providing a detailed analysis of audio signals over time.
- It is widely used in audio processing applications, such as speech recognition and music analysis, to capture the textural details of sounds effectively.

Spectrogram of speech signal

- A spectrogram of a speech signal displays the intensity of various frequencies over time, offering a visual snapshot of speech characteristics such as pitch, tone, and rhythm.

Source: Oppenheim, Alan V., and Ronald W. Schafer. *Discrete-Time Signal Processing*. 3rd ed., Prentice-Hall, 2009 [4]

Chroma

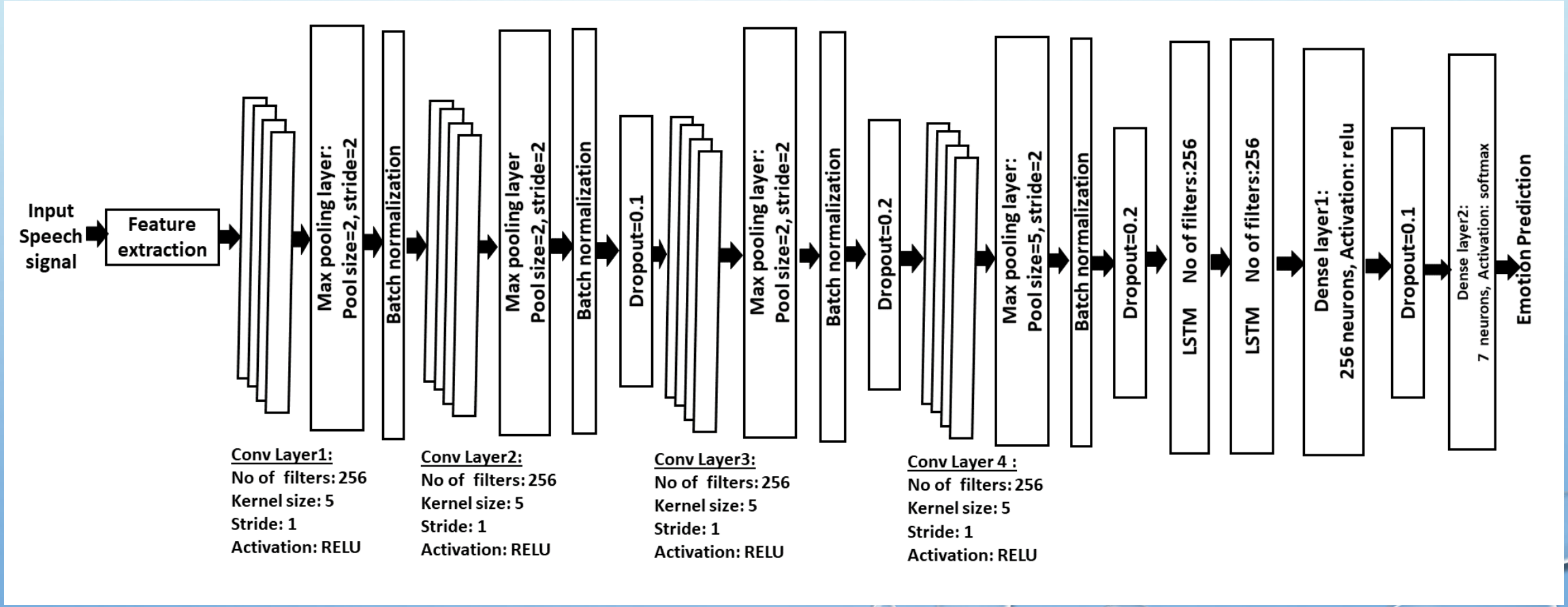
- Chroma features represent the pitch content of speech signals, highlighting harmonic patterns that are essential for speaker and speech recognition.

Tonnetz

- Tonnetz features capture harmonic relationships between audio frequencies.

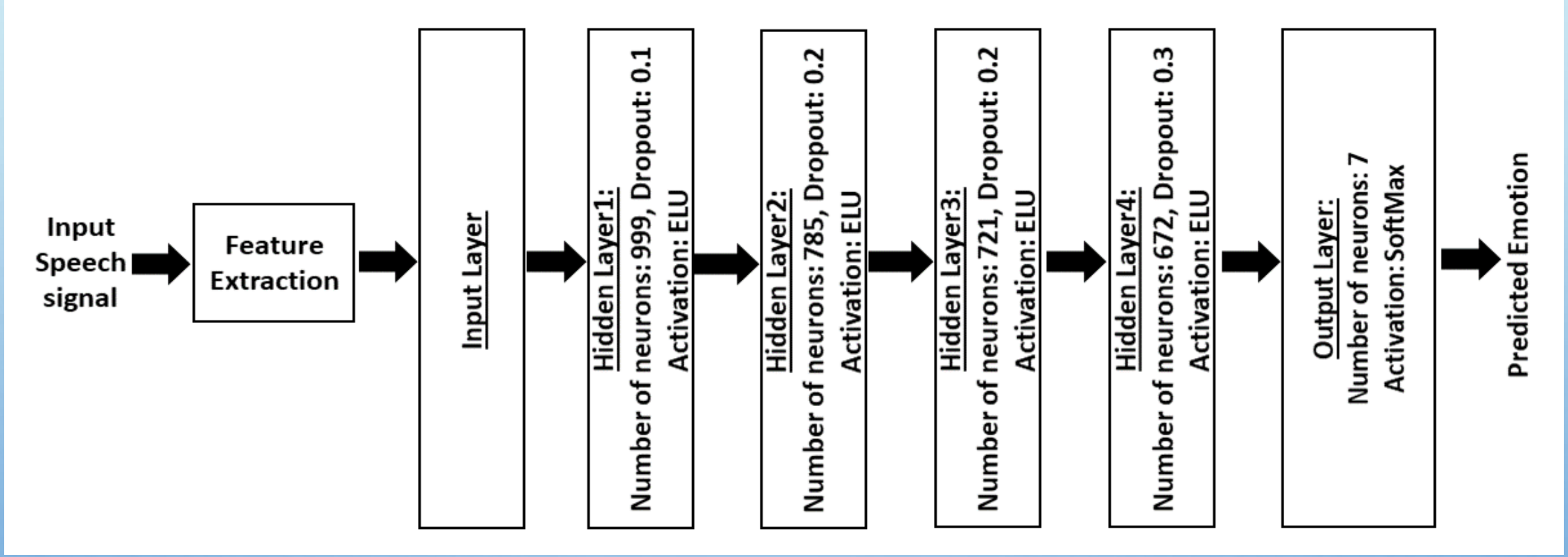
Classifiers

CNN-LSTM Architecture



Classifiers

DNN Architecture



5-Fold cross-validation Strategy

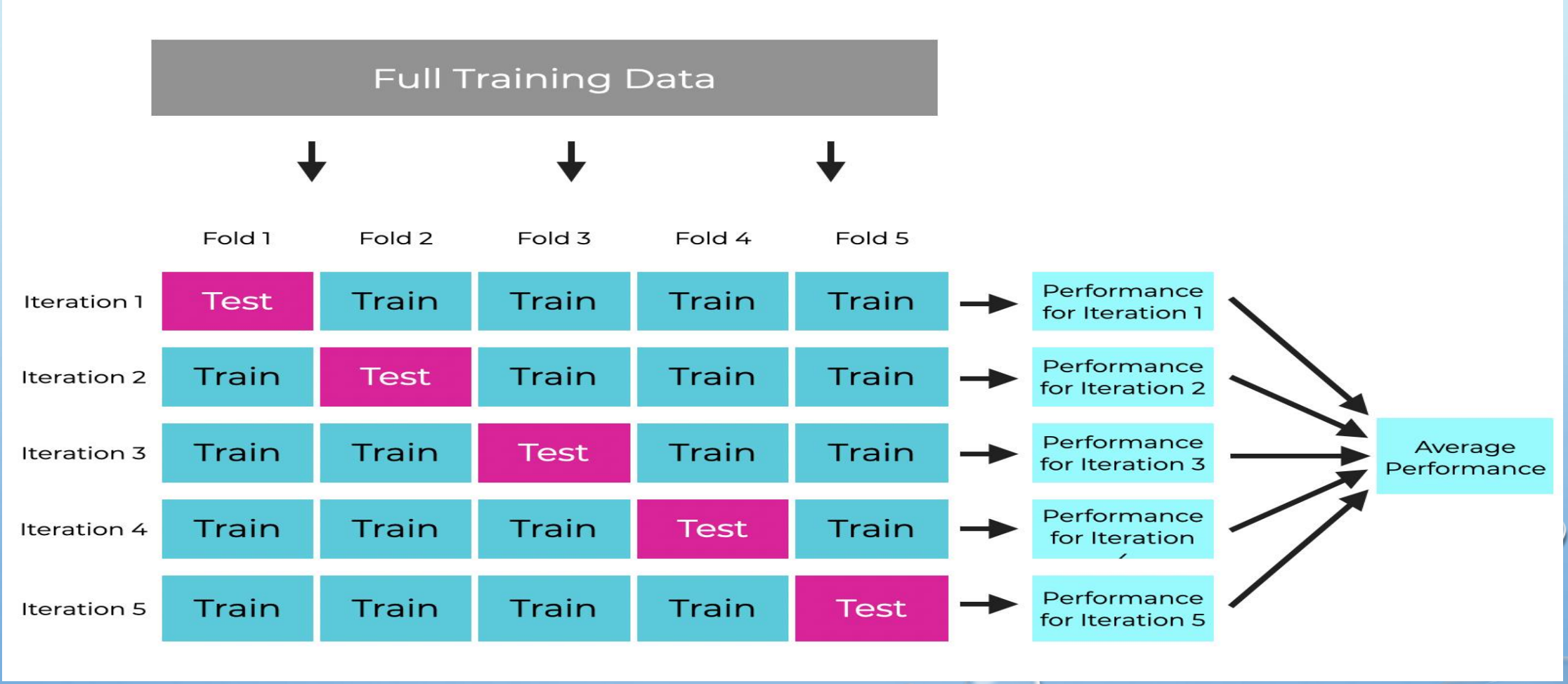
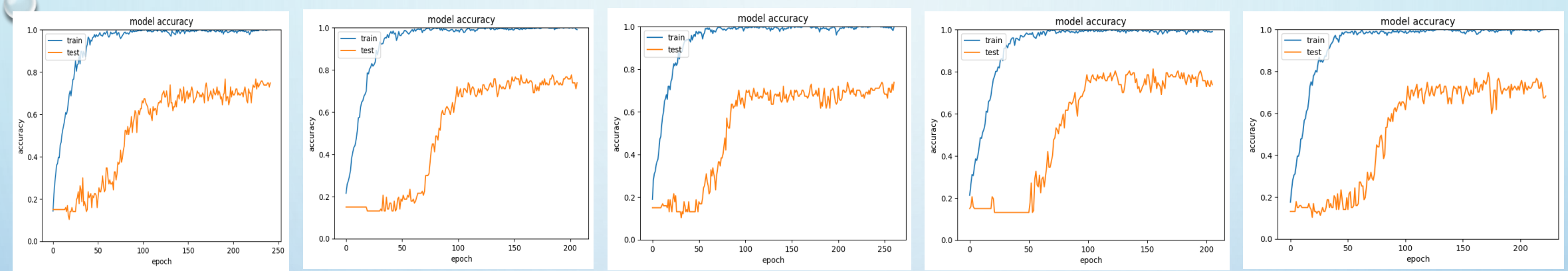


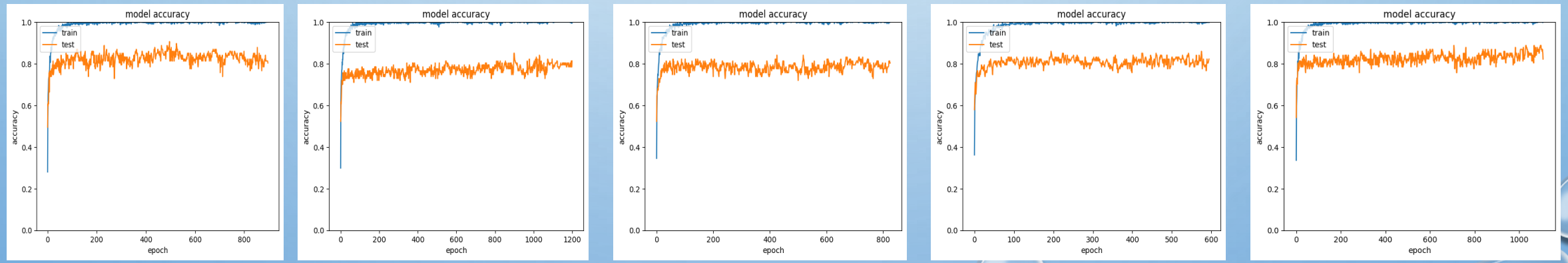
Figure-4 : 5-fold cross-validation diagram [5]

Results And Analysis

CNN+LSTM 5-Fold Accuracy Plots



DNN 5-Fold Accuracy Plots



Results And Analysis

Using 5-Fold cross-validation strategy

Fold	Validation accuracy with CNN+LSTM in %	Validation accuracy with DNN in %
Fold-1	76.63	90.65
Fold-2	75.70	85.05
Fold-3	73.83	84.98
Fold-4	81.38	85.98
Fold-5	79.44	88.79
Average of 5-Folds	77.38	87.09

Results And Analysis

Individual accuracies from train-test-split

Emotion	precision with CNN+LSTM model	precision with DNN model
anger	0.80	0.94
boredom	0.71	0.93
disgust	0.80	0.91
fear	0.80	0.82
happiness	0.71	0.80
neutral	0.73	0.79
sadness	0.92	0.86

Conclusion

- The DNN model achieved an impressive average accuracy of 87.09% across the five folds, outperforming the CNN+LSTM model, which achieved an average accuracy of 77.38%.
- A combination of features (Mel Spectrogram, MFCC, Delta MFCC, Tonnetz and Chroma) gives better accuracy than the individual.
- The impact of various hyperparameters and training strategies significantly influences model performance in optimizing accuracies for deep learning-based speech emotion recognition.
- DNN is very simple compared to the CNN+LSTM model and requires very little time to execute compared to CNN+LSTM.

References

[1] S. Mishra, P. Warule, and S. Deb, “Speech emotion classification using featureleveland classifier-level fusion,” Evolving Systems, vol. 15, pp. 1–14, 11 2023

[2] D. Bertero, F. Siddique, C.-S. Wu, Y. Wan, R. Chan, and P. Fung, “Real-timespeech emotion and sentiment recognition for interactive dialogue systems,” 012016, pp. 1042–1047

[3] A. Yadav and D. K. Vishwakarma, “A multilingual framework of cnn and bi-lstmfor emotion classification,” in 2020 11th international conference on computing,communication and networking technologies (ICCCNT). IEEE, 2020, pp. 1–6

[4] Oppenheim, Alan V., and Ronald W. Schafer. *Discrete-Time Signal Processing*. 3rd ed., Prentice-Hall, 2009

[5] “DenseNet architecture, note=image, url=https://www.sharpsightlabs.com/blog/crossvalidation-explained/,.”

Thank you!