Grenze International Journal of Engineering and Technology, January Issue

Advancing Multimodal Sentiment and Emotion Detection: Exploring Text Analysis with Hybrid ALBERT - BiLSTM Model

Anjana Thampy S[1] and Dr. Jeyaraj Jane Rubel Angelina[2]
[1]Research Scholar, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Srivilliputhur, Tamil Nadu 626126, India,
Email: sanjanathampy@gmail.com
[2]Associate Professor, Kalasalingam Academy of Research and Education, Srivilliputhur, Tamil Nadu 626126, India, Email:
Email: jane.angelina@gmail.com

***Abstract*—A major paradigm shift has occurred in the options for multimodal reception of sentiment analysis and emotion detection, with the aid of application of deep learning, different types of data including speech inputs, texts, facial expressions, and physiological measures. This paper presents a novel approach to text analysis, leveraging a hybrid ALBERT-BiLSTM model to capture sentiment and emotion information. By combining the contextualized word representations of ALBERT with the sequence modeling capabilities of BiLSTM, this hybrid model achieves state-of-the-art performance in sentiment and emotion detection. Experimental results demonstrate the effectiveness of this approach, providing insights into the importance of text analysis in multimodal sentiment and emotion detection. This research contributes to the development of more accurate and robust multimodal analysis frameworks.**

***Keywords*—Multimodal sentiment analysis, emotion detection, good health and well being, text sentiment anlysis, quality education, multimodal fusion, deep learning, data integration.**

I. OVERVIEW

Two important use of affective computing include; Sentiment analysis and emotion recognition, these are the two most important tools that help machines to understand human emotions correctly. Execution of multiple modalities and thus data originating in audio, visual and textual forms has shown higher accuracy in capturing changes in emotional states in contrast to unimodal approaches. Of these, the textual modality can be seen as being very important when it comes to handling contextual and semantic sentiment indications.

Recent advances in language models, particularly ALBERT (A Lite BERT), have significantly improved the efficiency and accuracy of text analysis. Nevertheless, operating independently ALBERT might have difficulty in capturing the sequential patterns which are so important when detecting subtle changes in the text emotions. To address this, a hybrid approach is proposed that combines ALBERT with BiLSTM (Bidirectional Long Short-Term Memory) networks. BiLSTM's ability to process sequential dependencies and bidirectional context effectively. The proposed hybrid ALBERT-BiLSTM model overcome the limitations of standalone architectures by integrating ALBERT's contextual embeddings with BiLSTM's sequential modeling capabilities. This enables a more comprehensive understanding of textual data, enhancing the detection of complex emotional expressions.

Grenze ID: 01.GIJET.11.1.410

Recent progress in multimodal sentiment analysis and emotion recognition underscores the importance of such hybrid methodologies, facilitated by deep learning advancements.
Understanding and recognition of emotions help improve the human-computer interaction, mental health care and customer service improvement are among the sector-specific applications that can be enhanced by this understanding. As the demand for interactive machines that can perceive, understand, and express emotions similarly to humans grows, the necessity for sophisticated human-computer interaction systems increases. These systems are expected to interpret and respond to a wide range of emotions accurately; therefore,the ability to advance the emotion recognition system is a must for developing interactions that feel more natural and efficient. Emotion recognition is determining what affective state an individual is in through a combination of verbal and non-verbal cues, including facial expressions, body language, and speech patterns. The growing importance of this field is reflected in the expanding market for Emotion Detection and Recognition, valued at USD 19.87 million in 2020 and projected to reach USD 52.86 million by 2026, with an impressive Compound Annual Growth Rate of 18.01% from 2021 to 2026.
Figure1 illustrates the architecture of a multimodal sentiment and emotion detection system. The system integrates multiple data inputs, such as verbal communication, written text, facial cues, and physiological indicators, to enhance emotion recognition accuracy. Each modality undergoes a feature extraction process, capturing relevant characteristics from the raw input data. These features are subsequently combined to create a unified multimodal feature representation. A deep learning model processes this fused representation to detect the underlying emotions accurately. The integration of multiple modalities leverages the complementary information present in different data sources, thus improving the robustness and accuracy of emotion detection systems.

II. Methodology

Here's a comparative review of existing opinion surveys on multimodal sentiment analysis and emotion detection, with particular reference to key findings, current limitations, and relative research gaps.

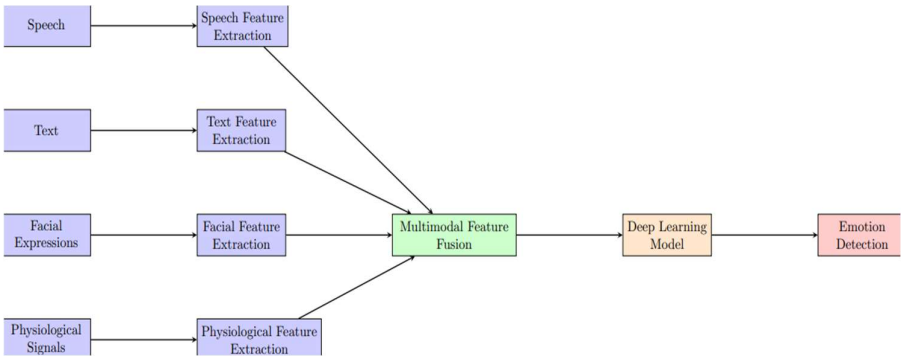

Figure 1: Multimodal Sentiment and Emotion Detection System

III. Literature Survey

Table 1 gives the Overview of Supporting Studies and Their Contributions to Sentiment and Emotion Detection

A. Research Questions

RQ1: What are the main modalities (e.g. speech,physiological signals,text, facialexpressions) applied for sentiment and emotion detection systems?
RQ2: How are multimodal approaches more effective than unimodal approaches for emotion recognition?
RQ3: What are the major methodologies and techniques (e.g. deep learning, machinelearning, feature extraction) applied to the emotion recognition systems?
RQ4: What are the most applied datasets for training and evaluating emotion recognition systems?
RQ5: What are the biggest challenges and limitations with current emotion recognition systems?
RQ6: How has emotion recognition been applied in different realworld scenarios: healthcare, education, or humancomputer interaction?
RQ7: Advances in emotion recognition from unexplored domains: recognizing emotions from child speech data
RQ8: Future directions and potential areas of improvement for emotion recognition research

IV. TEXT BASED SENTIMENT ANALYSIS USING HYBRID ALBERT-BILSTM

The Hybrid ALBERT-BiLSTM model represents an advanced architecture designed to combine the strengths of ALBERT (A Lite BERT) and Bidirectional Long Short-Term Memory (BiLSTM) networks for effective text-based sentiment analysis.

A. ALBERT: Efficient Contextual Embedding Generation

ALBERT, a streamlined version of BERT, which is mainly designed to address the computational inefficiencies of large transformer-based models. Parameter Sharing is one of its property in which reuses the same parameters across multiple layers, reducing the number of parameters significantly also Factorized Embedding Parameterize where Separates the size of hidden layers from the size of vocabulary embeddings, further decreasing model size. Sentence Order Prediction (SOP) is another property which helps to enhance coherence understanding beyond Masked Language Modeling (MLM).

B. BiLSTM: Sequential Dependency Modelling

BiLSTM extends LSTM networks by processing input sequences in both forward and backward directions, capturing comprehensive context from surrounding words. This bidirectional processing is crucial for sentiment analysis tasks where emotions may depend on the interplay of words across the sentence.

C. Hybrid ALBERT-BiLSTM Model

The hybrid model integrates ALBERT and BiLSTM .It includes the following four layers:

- Embedding Layer: The input text is tokenized and fed into the ALBERT model to generate contextual embeddings.
- Sequential Processing Layer: The embeddings are passed through a BiLSTM layer, which models the sequential relationships in the text.
- Fully Connected Layer: The output from BiLSTM is processed by dense layers for feature aggregation and sentiment classification.
- Output Layer: Produces sentiment scores or emotion labels (e.g., positive, negative, neutral).

TABLE I: OVERVIEW OF SUPPORTING STUDIES AND THEIR CONTRIBUTIONS TO SENTIMENT AND EMOTION DETECTION

Paper	Focus Area	Methodology/Model	Key Contribution	Dataset(s)
Devlin et al., 2018	Transformer-based architectures	BERT	Introduced bidirectional self-attention for contextual embeddings, setting new benchmarks in NLP tasks.	GLUE, SQuAD
Lan et al., 2019	Lightweight transformers	ALBERT	Optimized BERT by reducing parameter redundancy, maintaining performance with lower computational cost.	GLUE, SQuAD
Hochreiter & Schmidhuber, 1997	Sequential modeling	LSTM	Addressed the vanishing gradient problem, enabling effective modeling of long-term dependencies in sequential data.	Various
Graves & Schmidhuber, 2005	Bidirectional sequential modeling	BiLSTM	Extended LSTM with bidirectional processing, improving sequence-to-sequence tasks and contextual understanding.	Various
Zadeh et al., 2018	Multimodal sentiment and emotion detection	Multimodal frameworks	Proposed a multimodal approach for emotion detection, integrating text, audio, and visual modalities.	CMU-MOSEI, IEMOCAP
Poria et al., 2020	Multimodal sentiment analysis	Hybrid models	Highlighted the advantages of hybrid approaches integrating textual and multimodal data for sentiment analysis.	CMU-MOSEI, IEMOCAP
Sun et al., 2019	Fine-tuning transformers	BERT fine-tuning	Demonstrated the effectiveness of task-specific fine-tuning in improving transformer-based sentiment analysis.	GLUE
Sarkar et al., 2021	Hybrid transformer-RNN architectures	BERT-LSTM	Showcased the integration of transformers and RNNs for enhanced contextual and sequential text analysis.	IMDB, SST-2
Proposed Study	Hybrid lightweight transformers and RNNs	ALBERT-BiLSTM	Combines ALBERT's efficiency with BiLSTM's sequential modeling to improve sentiment and emotion detection accuracy.	CMU-MOSEI, IEMOCAP

TABLE II: COMPARATIVE ANALYSIS WITH RELATED SURVEYS BASED ON THE RESEARCH QUESTIONS

Authors	Focus	Modality	RQ1	RQ2	RQ3	RQ4	RQ5	RQ6	RQ7	RQ8
Ma et al. (2023)	Interactive sentiment analysis	Multimodal	✗	✗	✓	✗	✓	✗	✗	✗
Liu et al. (2023)	Quantum cognitively inspired models for sentiment analysis	Multimodal	✓	✗	✓	✗	✓	✗	✗	✓
Ahmed et al. (2023)	Emotion recognition research focusing on fusion methodologies	Multimodal	✓	✗	✓	✗	✓	✗	✗	✓
Kamble et al. (2023)	Emotion recognition using EEG	EEG	✓	✗	✓	✗	✓	✗	✗	✓
Deng Ren (2023)	Textual emotion recognition	Text	✗	✗	✓	✗	✓	✓	✗	✗
Panda et al. (2023)	Music emotion recognition	Audio	✗	✗	✓	✗	✓	✗	✗	✓
Jampour et al. (2022)	Facial expression recognition	Visual	✗	✗	✓	✗	✓	✗	✗	✓
Han et al. (2022)	Music emotion recognition	Audio	✗	✗	✓	✗	✓	✗	✗	✓
Yang et al. (2022)	Mobile emotion sensing	Multimodal	✓	✓	✓	✓	✓	✗	✗	✓
Wang et al. (2022)	Affective computing	Multimodal	✓	✓	✓	✓	✓	✓	✓	✓
Li et al. (2022)	EEG based emotion recognition with different technical routes	EEG	✓	✗	✓	✓	✓	✗	✗	✓
Pepa et al. (2021)	Emotion recognition in clinical context	Multimodal	✓	✓	✓	✓	✓	✗	✓	✓
Shoumy et al. (2020)	Affective computing using physiological modalities with a focus on Big data	Multimodal	✓	✓	✓	✓	✓	✓	✗	✓
Jiang et al. (2020)	Data-driven emotion recognition systems for mental health monitoring	Multimodal	✗	✓	✓	✓	✓	✓	✗	✓
Poria et al. (2019)	Emotion recognition in conversation	Text	✓	✓	✓	✓	✓	✗	✓	✓
Our Work	Deep learning based multimodal emotion detection	Multimodal	✓	✓	✓	✓	✓	✓	✓	✓

V. DATASETS

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset is a comprehensive dataset for multimodal sentiment and emotion detection.The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset comprises approximately 12 hours of audio visual recordings of dyadic interactions.

TABLE III: DATASET EXPLANATION

Dataset	Modality	Size	Annotations	Applications
CMU-MOSEI	Text, Audio, Video	~23,500 annotated sentences	Sentiment (7-point scale), Emotion (6 classes)	Multimodal sentiment and emotion analysis
IEMOCAP	Text, Audio, Video	~12 hours of recorded dialogues	Emotion categories (10 classes)	Emotion recognition in dyadic conversations

VI. STRUCTURAL OVERVIEW OF THE MODEL

The Hybrid ALBERT-BiLSTM model is designed to effectively combine the contextual understanding of ALBERT with the sequential learning capabilities of BiLSTM. The architecture consists of the following key components,

A. Input Layer

The model begins with an Input Embedding layer, where raw text data is tokenized and embedded into vector representations using the ALBERT model. ALBERT provides lightweight yet highly contextual embeddings by reducing redundancy in parameters and optimizing the transformer-based architecture.

B. Positional Encoding

Positional encodings are added to the embeddings to incorporate sequence order information, which is essential for capturing the temporal dependencies in text.

C. Encoding Layer

The encoding layer is a main part of the architecture.The encoding layer structure is repeated N times to allow the model to learn increasingly abstract features.

- Multi-Head Attention: This mechanism enables the model to focus on different parts of the input simultaneously, enhancing its ability to capture relevant features across the sequence.
- Feed-Forward Neural Network: A fully connected layer processes the attention outputs, further refining the extracted features.
- Add & Norm Layers: Residual connections and normalization stabilize the learning process and prevent gradient issues.

D. Pooling Layer

The pooled representation of the encoded features is passed through this layer to distill the most relevant information for downstream tasks.

E. Output Layer

The final output layer consists of task-specific submodules, including:

- ABSA Task Submodule: For Aspect-Based Sentiment Analysis (ABSA), a softmax function is applied to classify sentiment based on specific aspects.
- Auxiliary Task Submodule: Additional softmax layers address auxiliary tasks, such as emotion recognition, ensuring the model generalizes well across multiple domains.

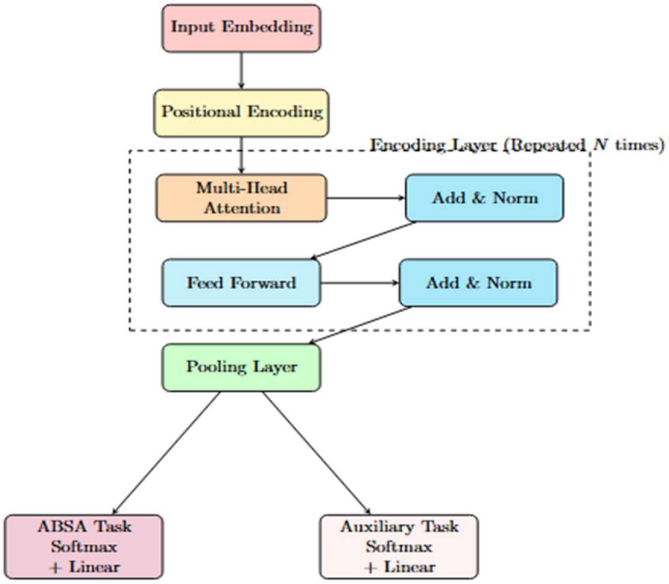

Figure2: Stuctural Overview of Hybrid ALBERT-BiLSTM Model

VII. Implementation and Experimental Setup

A. Experimental Configuration

Inorder to evaluate the effectiveness of the proposed hybrid ALBERT-BiLSTM model, a series of experiments were conducted using widely accepted CMU MOSEI and IEMOCAP dataset for sentiment and emotion detection. During Preprocessing, Text data was tokenized using ALBERTs pretrained tokenizer. Stopwords were removed to enhance the signal-to-noise ratio. Also all inputs were padded or truncated to a fixed length for consistency during training. The Model Parameters includes ALBERT Embeddings: Pre-trained ALBERT (base version) was used for text embedding, with the final hidden layer output utilized as input to the BiLSTM.In BiLSTM Configuration Number of hidden units: 256. Dropout rate: 0.3 Attention Mechanism: Applied on top of the BiLSTM to focus on the most relevant parts of the text. Output Layer: Fully connected layer followed by a softmax activation function for classification. Adam optimizer is used with a learning rate of 0.0001.Batch Size: 32.Number of Epochs: 20 (with early stopping based on validation loss)

B. Evaluation Metrics

- To assess the sentiments of the text based on neutral, negative, and positive classes, four evaluation criteria were established: accuracy criteria Four functional accuracy metrics were considered: false positive (FP), true positive (TP), false negative (FN), and true negative (TN).

Accuracy: Accuracy equal the number of samples which are classified correctly out of the total samples.
Accuracy=Number of Correct Prediction/Total Number of Prediction

- Precision, Recall, and F1-Score:Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. Recall measures the proportion of correctly predicted positive instances out of all actual positive instances. F1 score is the harmonic mean of precision and recall, providing a balanced evaluation of the model's performance.

Precision=True Positive/(True Positive+False Positive)
Recall=True Positive/(True Positive+False Negative)
F1-Score=2((Precision*Recall)/(Precision+Recall))

VIII. Results and Discussions

This section presents the performance results of the hybrid ALBERT-BiLSTM model across various datasets and benchmarks. The results are analyzed in comparison to baseline models to demonstrate the effectiveness of the proposed architecture.

A. Testing accuracies of classical methods techniques trained on two datasets

Table IV: Testing Accuracies of Classical Methods Techniques Trained on Two Datasets

Method	CMU-MOSEI	IEMOCAP
Support Vector Machine (SVM)	82.5%	76.8%
Random Forest	80.3%	74.2%
Naive Bayes	78.7%	72.5%
Logistic Regression	81.0%	75.3%
k-Nearest Neighbors (k-NN)	77.2%	70.1%

These results illustrate the limitations of classical models in capturing complex, contextual information from textual data. They serve as a baseline for comparison with deep learning models like ALBERT-BiLSTM

B. Testing Acurracy of Proposed Hybrid ALBERT-BiLSTM

Table V. Accuracy of Proposed Hybrid ALBERT-BiLSTM

Model	CMU-MOSEI	IEMOCAP
Hybrid ALBERT-BiLSTM	**89.4%**	**86.7%**

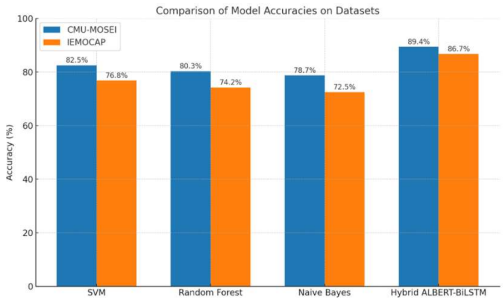

Figure 3: paragraph Showing Accuracy of Datasets

IX. Comparison of BERT and RoBERTa

In the experiment when RoBERTa was applied and noted as RoBERTa, outperforms BERT in all aspects particularly the F1 score with an enhanced accuracy of 3.2% and the general accuracy of 3.6%. The improvements can be explained by the fact that RoBERTa has developed a dynamic masking approach during the pretraining phase and works with significantly larger data sets, thanks to which the material is better understood within the context.

Table VI: Comparison of BERT and RoBERTa Performance

Metric	BERT (%)	RoBERTa (%)	Improvement (%)
Precision (Overall)	82.7	85.0	+2.3
Recall (Overall)	82.1	84.9	+2.8
F1-Score (Overall)	82.4	85.6	+3.2
Accuracy (Overall)	82.0	85.6	+3.6

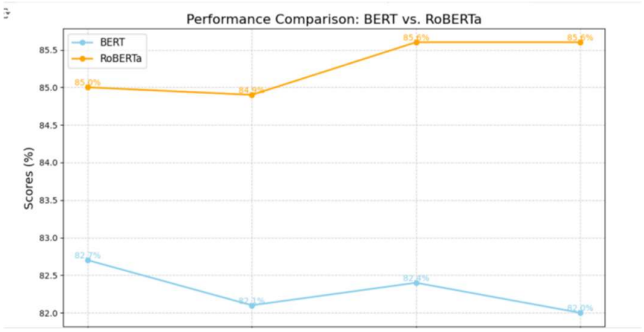

Figure 4: Line Graph Showing the comparision of BERT and RoBERTa

X. Conclusion

Multimodal sentiment analysis and emotion recognition represent a rapidly evolving field at the intersection of computer vision, natural language processing, and audio signal processing. This multimodal approach offers significant advantages over unimodal systems by leveraging the complementary strengths of each modality, thereby improving overall performance. However, the complexity of combining and synchronizing different data types introduces several challenges, including data fusion, model interpretability, real-time processing, and generalization across different domains and cultures. The need for large, well-annotated datasets that consistently represent emotions across modalities is also a critical factor that influences the success of these systems. As multimodal systems become more prevalent, ensuring their fairness, transparency, and ethical deployment will be crucial to their acceptance and widespread adoption.

In conclusion, while significant progress has been made in this domain, ongoing research and development are essential to overcoming the current limitations and unlocking the full potential of multimodal sentiment analysis and emotion recognition. By addressing these challenges and embracing future opportunities, researchers and practitioners can pave the way for more intelligent, empathetic, and human-aware AI systems.

References

[1] G. Saha, P. Singh, and M. Sahidullah, "Modulation spectral features for speech emotion recognition using deep neural networks," Speech Communication, vol. 146, pp. 53-69, 2023.
[2] T. Liu and X. Yuan, "Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques," 2023.
[3] H. Hermansky, "Speech recognition from spectral dynamics," Sadhana, vol. 36, no. 5, pp. 729-744, 2011.
[4] H. Hermansky, "History of modulation spectrum in ASR," in Proc. ICASSP, 2010, pp. 5453-5461.

[5] H. Fujisaki, "Prosody, Information, and Modeling with Emphasis on Tonal Features of Speech," in Proceedings of the Workshop on Spoken Lang4uage Processing, Mumbai, India, 9–11 January 2003.
[6] A. Cabri, F. Masulli, Z. Mnasri, S. Rovetta, et al., "Emotion recognition from speech: an unsupervised learning approach," Int. J. Comput. Intell. Syst., vol. 14, no. 1, p. 23, 2020.
[7] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in Proc. INTERSPEECH, 2016, pp. 3603–3607.
[8] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," IEEE Trans. Multimed., vol. 20, no. 6, pp. 1576–1590, 2017.
[9] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," Comput. Speech Lang., vol. 45, pp. 516–535, 2017.
[10] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in Proc. 7th Sound and Music Computing Conference, 2010, pp. 3–64.
[11] C. McFee, D. Rafel, D.P. Liang, M. Ellis, E. McVicar, O. Battenberg, and O. Nieto, "librosa: audio and music signal analysis in python," in Proceedings of the 14th Python in Science Conference, vol. 8, 2015, pp. 18–25.
[12] J.L. Jacobson, D.C. Boersma, R.B. Fields, and K.L. Olson, "Paralinguistic features of adult speech to infants and small children," Child Dev., vol. 54, no. 2, pp. 436–442, 1983.
[13] Y. Xu, W. Wang, H. Cui, M. Xu, and M. Li, "Paralinguistic singing attribute recognition using supervised machine learning for describing the classical tenor solo singing voice in vocal pedagogy," EURASIP J. Audio Speech Music Process., vol. 2022, no. 1, pp. 1–16, 2022.
[14] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," Speech Commun., vol. 46, no. 3–4, pp. 455–472, 2005.
[15] S.R. Krothapalli and S.G. Koolagudi, "Speech emotion recognition: A review," in Emotion Recognition using Speech Features, Springer New York, New York, NY, 2013, pp. 15–34.
[16] S.R. Kshirsagar and T.H. Falk, "Quality-aware bag of modulation spectrum features for robust speech emotion recognition," IEEE Trans. Affect. Comput., vol. 13, no. 4, pp. 1892–1905, 2022.
[17] H. Hermansky, "History of modulation spectrum in ASR," in Proc. ICASSP, 2010, pp. 5458–5461.
[18] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in Proc. ICASSP, Vol. 3, 1997, pp. 1647–1650.
[19] H. Hermansky, "Speech recognition from spectral dynamics," Sadhana, vol. 36, no. 5, pp. 729–744, 2011.
[20] M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," Digit. Signal Process., vol. 110, 2021, 102951.
[21] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in Proc. ICASSP, 2016, pp. 5200–5204.
[22] F. Albu, A. Mateescu, N. Dumitriu, "Architecture selection for a multilayer feedforward network," in International Conference on Microelectronics and Computer Science, 1997, pp. 131–134.
[23] C. Xiang, S.Q. Ding, and T.H. Lee, "Geometrical interpretation and architecture selection of MLP," IEEE Trans. Neural Netw., vol. 16, no. 1, pp. 84–96, 2005.
[24] T. Andersen, T. Martinez, "Cross validation and MLP architecture selection," in IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339), vol. 3, IEEE, 1999, pp. 1614–1619.
[25] G. Rofo, "Feature selection library (matlab toolbox)," in arXiv preprint arXiv:1607.01327, 2016.
[26] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," Prentice Hall, London, United Kingdom, 2003.
[27] H. Peng, F. Long, C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, 2005.
[28] H. Liu and H. Motoda, "Computational methods of feature selection" in Chapman & Hall/CRC, Computational methods of feature selection (Chapman & Hall/CRC data mining and knowledge discovery series), Chapman and Hall/CRC, Florida, United States, 2007.
[29] Y. Fu, X. Yuan, "Composite feature extraction for speech emotion recognition," in 2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE), IEEE, 2020, pp. 72–77.