REFERENCES

[1] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[2] X. Wang and L. Xu, "Speech perception in noise: Masking and unmasking," *Journal of Otology*, vol. 16, no. 2, pp. 109–119, 2021.

[3] J. Pons, O. Slizovskaia, R. Gong, E. Gomez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," pp. 2744–2748, 2017.

[4] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent systems with applications*, vol. 16, p. 200115, 2022.

[5] C. Mouton, J. C. Myburgh, and M. H. Davel, "Stride and translation invariance in cnns," in *Southern African Conference for Artificial Intelligence Research*. Springer, 2020, pp. 267–281.

[6] F. Bieder, R. Sandkühler, and P. C. Cattin, "Comparison of methods generalizing max-and average-pooling," *arXiv e-prints*, pp. arXiv–2103, 2021.

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems," *arXiv preprint arXiv:1904.03476*, 2019.

[8] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[9] H. Zhao, Z. Gao, Y. Wang, R. Xiong, and Y. Zhang, "Adaptive wavelet-positional encoding for high-frequency information learning in implicit neural representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 430–10 438.

[10] V. Passricha and R. K. Aggarwal, "A comparative analysis of pooling strategies for convolutional neural network based hindi asr," *Journal of ambient intelligence and humanized computing*, vol. 11, no. 2, pp. 675–691, 2020.

[11] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," pp. 464–472, 2016.

[12] R. Huang and Y. Xie, "A cnn-based multi-scale pooling strategy for acoustic scene classification," *IEICE TRANSACTIONS on Information and Systems*, vol. 107, no. 1, pp. 153–156, 2024.

[13] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," pp. 530–546, 2014.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *International conference on learning representations*, 2018.

[16] O. Rippel, J. Snoek, and R. P. Adams, "Spectral representations for convolutional neural networks," *Advances in neural information processing systems*, vol. 28, 2015.

[17] T. Rathi and M. Tripathy, "Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review," *Speech Communication*, p. 103102, 2024.

[18] Z. Liu, "Ai-driven classification and trend analysis of piano music genres using large language models," *International Journal of Information and Communication Technology*, vol. 26, no. 12, pp. 32–48, 2025.

[19] Y. Xin, D. Yang, and Y. Zou, "Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss," pp. 1–5, 2023.

[20] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," vol. 1, pp. 113–116, 2002.

[21] E. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*. London, UK: Academic Press, 1999.

[22] G. Mohmad and R. Delhibabu, "Speech databases, speech features and classifiers in speech emotion recognition: A review," *IEEE Access*, 2024.

[23] R. Nirthika, S. Manivannan, A. Ramanan, and R. Wang, "Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study," *Neural Computing and Applications*, vol. 34, no. 7, pp. 5321–5347, 2022.

[24] G. Strang and T. Nguyen, *Wavelets and filter banks*. SIAM, 1996.

[25] H. R. Seresht and K. Mohammadi, "Environmental sound classification with low-complexity convolutional neural network empowered by sparse salient region pooling," *IEEE Access*, vol. 11, pp. 849–862, 2022.

[26] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," pp. 530–546, 2014.

[27] V. Lyberatos, S. Kantarelis, E. Dervakos, and G. Stamou, "Challenges and perspectives in interpretable music auto-tagging using perceptual features," *IEEE Access*, 2025.

[28] O. K. Toffa and M. Mignotte, "Environmental sound classification using local binary pattern and audio features collaboration," *IEEE Transactions on Multimedia*, vol. 23, pp. 3978–3985, 2020.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[30] A. Lependin, P. Ladygin, V. Karev, and A. Mansurov, "Fourier chromagrams for fingerprinting, verification and authentication of digital audio recordings," pp. 263–275, 2023.

[31] C. Jones, A. Smith, and E. Roberts, "A sample paper in conference proceedings," in *Proc. ICASSP*, vol. II, Apr. 2003, pp. 803–806.

[32] O. A. Onasoga, N. Yusof, and N. H. Harun, "Audio classification-feature dimensional analysis," pp. 775–788, 2021.

[33] A. Natsiou and S. O'Leary, "Audio representations for deep learning in sound synthesis: A review," pp. 1–8, 2021.

[34] I. Bozhilov, R. Petkova, K. Tonchev, and A. Manolova, "A systematic survey into compression algorithms for three-dimensional content," *IEEE Access*, 2024.

[35] C. Huang, W. Talbott, N. Jaitly, and J. M. Susskind, "Efficient representation learning via adaptive context pooling," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9346–9355.

[36] Z. Liu, T. Shao, and X. Zhang, "Bcg signal analysis based on improved vmd algorithm," *Measurement*, vol. 231, p. 114631, 2024.

[37] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent systems with applications*, vol. 16, p. 200115, 2022.

[38] S. Yadav, D. M. Yadav, and K. R. Desai, "A comprehensive survey of automatic dysarthric speech recognition," *Int J Inf & Commun Technol ISSN*, vol. 2252, no. 8776, p. 8776.

[39] X. Lu, P. Shen, S. Li, Y. Tsao, and H. Kawai, "Temporal attentive pooling for acoustic event detection." pp. 1354–1357, 2018.

[40] K. Zhang, Z. Wu, J. Jia, H. Meng, and B. Song, "Query-by-example spoken term detection using attentive pooling networks," pp. 1267–1272, 2019.

[41] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognition Letters*, vol. 146, pp. 1–7, 2021.

[42] J. Zhang, L. Gao, B. Hao, H. Huang, J. Song, and H. Shen, "From global to local: Multi-scale out-of-distribution detection," *IEEE Transactions on Image Processing*, 2023.

[43] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[44] P. Swietojanski and S. Renals, "Differentiable pooling for unsupervised acoustic model adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1773–1784, 2016.

[45] A. Nfissi, W. Bouachir, N. Bouguila, and B. Mishara, "Sigwavnet: Learning multiresolution signal wavelet network for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2025.

[46] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. Moore, "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends in Hearing*, vol. 27, p. 23312165231209913, 2023.

[47] C.-C. Kao, B. Shi, M. Sun, and C. Wang, "A joint framework for audio tagging and weakly supervised acoustic event detection using densenet with global average pooling," *arXiv preprint arXiv:2008.03350*, 2020.

[48] M. Wolter, S. Lin, and A. Yao, "Neural network compression via learnable wavelet transforms," in *Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part II 29*. Springer, 2020, pp. 39–51.

[49] M. Rouvier, P.-M. Bousquet, and J. Duret, "Study on the temporal pooling used in deep neural networks for speaker verification," pp. 501–505, 2021.

[50] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," *arXiv preprint arXiv:1604.06338*, 2016.

[51] P. Lopez-Meyer, J. A. del Hoyo Ontiveros, H. Lu, and G. Stemmer, "Efficient end-to-end audio embeddings generation for audio classification on target applications," pp. 601–605, 2021.

[52] A. Smith, C. Jones, and E. Roberts, “A sample paper in journals,” *IEEE Trans. Signal Process.*, vol. 62, pp. 291–294, Jan. 2000.

[53] X. Wei, Y. Zhang, Y. Gong, and N. Zheng, “Kernelized subspace pooling for deep local descriptors,” pp. 1867–1875, 2018.

[54] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, “Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection,” pp. 646–650, 2022.

[55] M. Wolter and J. Garcke, “Adaptive wavelet pooling for convolutional neural networks,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1936–1944.

[56] G. Xu, W. Liao, X. Zhang, C. Li, X. He, and X. Wu, “Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation,” *Pattern recognition*, vol. 143, p. 109819, 2023.

[57] P. Liu, H. Zhang, W. Lian, and W. Zuo, “Multi-level wavelet convolutional neural networks,” *IEEE Access*, vol. 7, pp. 74 973–74 985, 2019.

[58] L. Zhao and Z. Zhang, “A improved pooling method for convolutional neural networks,” *Scientific Reports*, vol. 14, no. 1, p. 1589, 2024.

[59] K. C. Vidyasagar, K. R. Kumar, G. A. Sai, M. Ruchita, and M. J. Saikia, “Signal to image conversion and convolutional neural networks for physiological signal processing: A review,” *IEEE Access*, 2024.

[60] A. Vyas, S. Yu, and J. Paik, *Multiscale transforms with application to image processing*. Springer, 2018, vol. 1.

[61] I. Syafalni, C. Amadeus, N. Sutisna, and T. Adiono, “Efficient real-time smart keyword spotting using spectrogram-based hybrid cnn-lstm for edge system,” *IEEE Access*, vol. 12, pp. 43 109–43 125, 2024.

[62] P. Warden, “Speech commands: A dataset for limited-vocabulary speech recognition,” *arXiv preprint arXiv:1804.03209*, 2018.

[63] A. S. Abdulaziz, A. Dawood, and A. Daood, “Speaker identification and verification using convolutional neural network cnn,” *Tikrit Journal of Engineering Sciences*, vol. 32, no. 2, pp. 1–13, 2025.

[64] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” *arXiv preprint arXiv:1706.08612*, 2017.

[65] H. Yu, C. Liu, L. Zhang, C. Wu, G. Liang, J. Escorcia-Gutierrez, and O. A. Ghoneim, “An intent classification method for questions in" treatise on febrile

diseases" based on tinybert-cnn fusion model," *Computers in Biology and Medicine*, vol. 162, p. 107075, 2023.

[66] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[67] J. M. Oh, J. K. Kim, and J. Y. Kim, "Multi-detection-based speech emotion recognition using autoencoder in mobility service environment," *Electronics*, vol. 14, no. 10, p. 1915, 2025.

[68] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[69] P. Cherukuru and M. B. Mustafa, "Cnn-based noise reduction for multi-channel speech enhancement system with discrete wavelet transform (dwt) preprocessing," *PeerJ Computer Science*, vol. 10, p. e1901, 2024.

[70] E. Vincent, S. Watanabe, J. Barker, and R. Marxer, "The 4th chime speech separation and recognition challenge," *URL: http://spandh. dcs. shef. ac. uk/chime_challenge/(last accessed on 1 August, 2018)*, 2016.

[71] X. Zhao and O. Tsimhoni, "Real-time pitch/f0 detection using spectrogram images and convolutional neural networks," *arXiv preprint arXiv:2504.06165*, 2025.

[72] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *Ismir*, vol. 14, 2014, pp. 155–160.

[73] R. Chen, A. Ghobakhlou, and A. Narayanan, "Interpreting cnn models for musical instrument recognition using multi-spectrogram heatmap analysis: a preliminary study," *Frontiers in Artificial Intelligence*, vol. 7, p. 1499913, 2024.

[74] J. J. Bosch, F. Fuhrmann, and P. Herrera, "Irmas: a dataset for instrument recognition in musical audio signals," 2014.

[75] J. K. Bhatia, R. D. Singh, and S. Kumar, "Music genre classification," pp. 1–4, 2021.

[76] B. L. Sturm, "The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[77] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied soft computing*, vol. 52, pp. 28–38, 2017.

[78] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[79] A. Bansal and N. K. Garg, "Robust technique for environmental sound classification using convolutional recurrent neural network," *Multimedia Tools and Applications*, vol. 83, no. 18, pp. 54 755–54 772, 2024.

[80] K. J. Piczak, "Esc: Dataset for environmental sound classification," pp. 1015–1018, 2015.

[81] J. Surendiran, P. E. Prabhakar, M. M. Ibrahim, G. Saritha, V. B. Vijayan *et al.*, "A systemic review on automatic acoustic scene classification," in *2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*. IEEE, 2024, pp. 1–6.

[82] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-workshop on detection and classification of acoustic scenes and events*, 2017.

[83] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the dcase 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

[84] S. S. Ghintab and M. Y. Hassan, "Cnn-based visual localization for autonomous vehicles under different weather conditions," *Eng. Technol. J*, vol. 41, no. 02, pp. 375–386, 2023.

[85] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[86] A. Solanki and S. Pandey, "Music instrument recognition using deep convolutional neural networks," *International Journal of Information Technology*, vol. 14, no. 3, pp. 1659–1668, 2022.

[87] C. M. Eckhardt, S. J. Madjarova, R. J. Williams, M. Ollivier, J. Karlsson, A. Pareek, and B. U. Nwachukwu, "Unsupervised machine learning methods and emerging applications in healthcare," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 31, no. 2, pp. 376–381, 2023.

[88] S. A. Sepúlveda-Fontaine and J. M. Amigó, "Applications of entropy in data analysis and machine learning: A review," *Entropy*, vol. 26, no. 12, p. 1126, 2024.

[89] L. Tang, H. Tian, H. Huang, S. Shi, and Q. Ji, "A survey of mechanical fault diagnosis based on audio signal analysis," *Measurement*, p. 113294, 2023.

[90] S. Yadav, D. M. Yadav, and K. R. Desai, "A comprehensive survey of automatic dysarthric speech recognition," *Int J Inf & Commun Technol ISSN*, vol. 2252, no. 8776, p. 8776.

[91] S. Tanberk, V. Dağlı, and M. K. Gürkan, "Deep learning for videoconferencing: A brief examination of speech to text and speech synthesis," pp. 506–511, 2021.

[92] S.-W. Park, J.-S. Ko, J.-H. Huh, and J.-C. Kim, "Review on generative adversarial networks: focusing on computer vision and its applications," *Electronics*, vol. 10, no. 10, p. 1216, 2021.

[93] A. Klapuri and M. Davy, "Signal processing methods for music transcription," 2007.