



Exploring the Power of Empirical Mode Decomposition for Sensing the Sound of Silence: A Pilot Study on Mice Autism Detection via Ultrasonic Vocalisation

Chenhao Wu^{1,#}, Xiangjun Cai^{2,#}, Haojie Zhang^{3,4}, Tianrui Jia^{3,4},
Yilu Deng^{5,*}, Kun Qian^{3,4,*}, Björn W. Schuller^{6,7}, Yoshiharu Yamamoto⁸, and Jiang Liu^{1,*}

¹Graduate School of Fundamental Science and Engineering, Waseda University, Japan

²Faculty of Innovation Engineering, Macau University of Science and Engineering, China SAR

³Key Laboratory of Brain Health Intelligent Evaluation and Intervention (Beijing Institute of Technology), Ministry of Education, Beijing, China

⁴School of Medical Technology, Beijing Institute of Technology, Beijing, China

⁵School of Economics and Management, North China Electric Power University, Beijing, China

⁶GLAM – the Group on Language, Audio, & Music, Imperial College London, UK

⁷CHI – Chair of Health Informatics, Technical University of Munich, Germany

⁸Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Japan

dengyilu@ncepu.edu.cn, qian@bit.edu.cn, jiang@waseda.jp

Abstract

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder, and mice models have become essential for studying its genetic and behavioural aspects. Ultrasonic Vocalisations (USVs) emitted by mice provide a promising biomarker for ASD detection, but existing methods relying on spectrogram-based features struggle to capture the complex, non-stationary, and multi-scale nature of USVs. To address this, we propose a novel multi-branch fusion model that integrates spectrogram-based features with multi-scale features extracted using Empirical Mode Decomposition (EMD), which decomposes USVs into Intrinsic Mode Functions (IMFs) to represent their inherent complexity better. Through systematic occlusion experiments, we identify high-frequency components, particularly IMF1, as critical for accurate ASD detection, highlighting the diagnostic relevance of high-frequency USV patterns. Our model achieves an Unweighted Average Recall (UAR) of 0.75 in subject-level classification, significantly outperforming existing methods. These findings provide valuable insights into the importance of multi-scale feature extraction and offer a robust framework for improving ASD diagnostics and research.

Index Terms: autism spectrum disorder, ultrasound vocalisations, empirical mode decomposition, multi-branch fusion model

1. Introduction

Autism Spectrum Disorder (ASD) is a multifaceted neurodevelopmental disorder characterised by significant impairments in social interaction, communication, and behaviour [1, 2, 3, 4, 5, 6]. While research on humans has provided valuable insights into the genetic and neurological foundations of ASD, animal models, particularly mice, have become indispensable for advancing our understanding of the disorder [7, 8]. Owing to the shared genomic regions between mice and humans, mice serve as an excellent model for studying ASD. For instance, mice with genetic mutations such as Shank3 or CNTNAP2, exhibit social and communicative behaviours that closely resemble those observed in humans with ASD [9, 10]. Analysis of ultrasonic vocalisations (USVs) in mice has emerged as a promising approach, as studies have demonstrated significant differences between wild-type (WT) and ASD model mice [11].

Research on human speech has revealed distinct prosodic differences between individuals with ASD and those without the condition. For example, individuals with ASD tend to speak at a slower rate [12] and display unique melodic patterns in their intonation [13]. These consistent patterns have prompted the development of machine learning techniques to detect ASD from speech data. Early methods employed hand-engineered features, such as the eGeMAPS feature set [14], combined with Support Vector Machines (SVMs) to differentiate between the speech of individuals with ASD and those with typical development, achieving over 75% Unweighted Average Recall (UAR) [15, 16] in binary classification tasks [17]. More recent approaches have utilised deep learning models, including Convolutional Neural Networks (CNNs) and fine-tuned Wav2Vec 2.0 architectures, to identify ASD from self-recorded speech samples. These methods have shown considerable improvements over traditional techniques [18].

In the realm of mice models, the study of USVs has emerged as a powerful tool for ASD research. Although mice produce sounds within the human-audible range, their ultrasonic vocalisations are more prevalent and have been shown to encode valuable information about various characteristics, including sex, age, and health status [19]. Recent advancements have employed deep neural networks to classify mice USVs, achieving high accuracy in tasks such as sex determination [20] and the classification of expert-defined vocalisation types [21]. Despite these successes, the use of machine learning to detect ASD in mice based on USVs remains an underexplored area. A pioneering study by Qian *et al.* [22] achieved a UAR of 66.6% in distinguishing ASD model mice from wild-type mice using a large-scale pre-trained audio neural network, marking the first successful application for ASD detection in mice.

Despite these advancements, existing methods for ASD detection in mice encounter several challenges. Traditional approaches predominantly rely on spectrogram-based features. Although these features effectively capture the time–frequency characteristics of USVs, they may not fully encapsulate the complex, multi-scale nature of these signals. USVs are inherently non-stationary and exhibit multi-scale characteristics that make them difficult to analyse using conventional methods. Additionally, the presence of noise and overlapping vocalisations further complicates the classification task. To address

these challenges, we propose a novel multi-branch fusion model that combines spectrogram-based features with multi-scale features extracted using Empirical Mode Decomposition (EMD). EMD is particularly well-suited for analysing non-stationary signals, as it decomposes the signal into multiple Intrinsic Mode Functions (IMFs) that represent different frequency components [23, 24, 25]. By integrating these IMF-based features with spectrogram features through a fusion network, our model achieves a more comprehensive representation of the signal, leading to improved classification performance. The contributions of this paper are threefold:

1. A novel multi-branch fusion model integrating spectrogram-based features with EMD-derived multi-scale representations for ASD detection in mice is proposed, with a subject-level UAR of 0.75 being achieved.
2. Insights are provided into the importance of multi-scale feature extraction in improving the performance of ASD detection models.
3. It is demonstrated that high-frequency components, particularly IMF1, are pivotal for ASD detection, as their removal leads to a marked degradation in model performance. This finding underscores the diagnostic relevance of high-frequency USV patterns in distinguishing ASD-related behaviours.

2. Proposed Method

In this section, we present our multi-branch fusion framework for ASD detection in mice, shown in Figure 1. The framework comprises four components: (1) Raw audio recordings are segmented into short clips, each capturing a distinct portion of a mouse’s vocalisation. This segmentation ensures independent treatment of each clip during both feature extraction and classification. (2) Each clip is converted into a time-frequency representation using the Short-Time Fourier Transform (STFT). A two-dimensional convolutional neural network (2D-CNN) then extracts spectral-temporal features that capture patterns, such as harmonic sweeps and broadband bursts, characteristic of ASD-specific USVs. (3) In parallel, each clip is adaptively decomposed into IMFs, with each IMF representing oscillatory modes at different time scales. These IMFs serve as multi-channel inputs to a one-dimensional convolutional neural network (1D-CNN), capturing fine-grained amplitude variations indicative of ASD-related vocal patterns. (4) Feature Fusion and Classification: The feature vectors from both branches are concatenated to form a unified representation, which is then processed by a fully connected (FC) layer for binary classification. A majority voting mechanism is subsequently applied at the subject-level to aggregate segment-level decisions, yielding a robust classification for each mouse.

2.1. Spectrogram Branch

A critical aspect of analysing mice USVs is the accurate capture of their extensive time–frequency structure, which includes harmonic relationships, spectral energy distributions, and transient broadband bursts. Traditional time–frequency analysis methods struggle with these dynamic, non-stationary characteristics. To overcome this, we introduce the Spectrogram Branch, which leverages the STFT to generate a two-dimensional frequency–time representation. This transformation provides a global perspective on spectral energy distribution, thereby offering insights into the evolution of sound patterns.

To enhance robustness, we emphasise the STFT magnitude

and apply logarithmic scaling, which stabilises amplitude variations and accentuates spectral envelopes. This preprocessing mitigates the adverse effects of abrupt intensity fluctuations, thereby facilitating improved feature learning for classification.

The resulting spectrograms are then processed by a 2D-CNN, whose two-dimensional receptive field efficiently identifies harmonic sweeps, frequency modulations, and broadband bursts. These are potential signatures of ASD-related vocal patterns. Through pooling layers, the CNN distils the spectrogram into a compact feature vector that preserves global spectral cues essential for distinguishing ASD-related from wild-type vocalisations.

By integrating this global time–frequency representation, the Spectrogram Branch forms the foundation for capturing large-scale acoustic structures, thereby complementing the multi-scale local features extracted by the subsequent EMD Branch.

2.2. EMD Branch

Despite the efficacy of spectrogram-based representations, mice USVs exhibit rapid transitions and multi-scale patterns that may not be fully discernible in the time–frequency domain alone. Transient oscillations, intertwined with slower periodic structures, often remain partially masked in spectrograms. To address these challenges, we introduce the EMD Branch, which adaptively decomposes each audio segment into a set of IMFs, each representing an oscillatory mode at a distinct time scale. Given an audio signal $x(t)$, its Empirical Mode Decomposition (EMD) is formally expressed as:

$$x(t) = \sum_{i=1}^N c_i(t) + r_N(t), \quad (1)$$

where $c_i(t)$ denotes the i th IMF and $r_N(t)$ represents the residual after extracting N modes. The extraction of each IMF is achieved via an iterative sifting process, which is formalised as:

$$c_i(t) = \lim_{k \rightarrow \infty} \left\{ \left[\prod_{j=1}^k (\mathcal{I} - \mathcal{S}) \right] r_{i-1}(t) \right\}, \quad r_0(t) = x(t), \quad (2)$$

$$r_i(t) = r_{i-1}(t) - c_i(t), \quad i = 1, 2, \dots, N, \quad (3)$$

where, the operator \mathcal{I} denotes the identity operator, which returns the input signal unaltered, while \mathcal{S} is the sifting operator that computes the local mean by averaging the upper and lower envelopes obtained via spline interpolation of the local extrema. The advantages of EMD-based decomposition are threefold:

Capturing local amplitude variations, since each IMF isolates localised oscillations, this branch is highly effective in detecting subtle amplitude modulations, which can serve as key indicators of ASD-related vocalisation patterns.

Reducing feature overlap, by decomposing the original signal into multiple IMFs, overlapping frequency components are disentangled, allowing the convolutional network to extract more discriminative features without interference from other scales.

Enhancing interpretability, each IMF corresponds to a distinct oscillatory mode, offering insight into how specific frequency bands contribute to ASD and wild-type differences. This granular decomposition aids post-hoc analysis, enabling a biologically meaningful understanding of ASD-related vocalisation characteristics.

To fully exploit these benefits, we feed the extracted IMFs into a 1D-CNN, designed to process multi-channel oscillatory

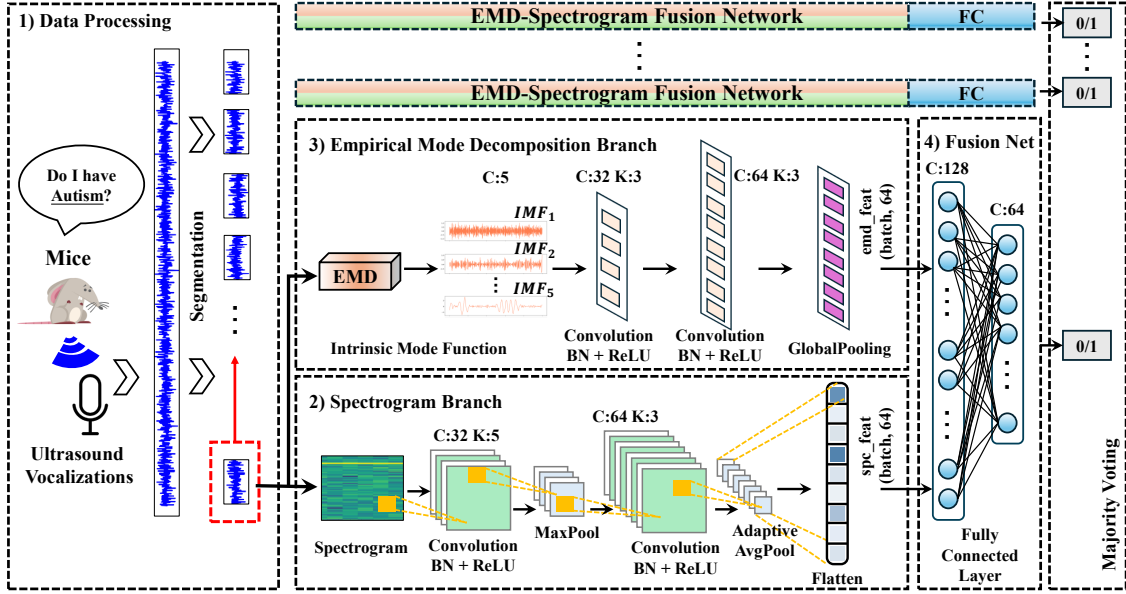


Figure 1: Overview of the proposed EMD-Spectrogram Fusion Network for ASD detection. (1) Data Processing: The raw audio signals are segmented into short clips to facilitate feature extraction. (2) Spectrogram Branch: Each clip is transformed into a spectrogram, from which time-frequency features are extracted using a 2D-CNN. (3) EMD Branch: EMD is applied to decompose each segment into IMFs, which are then processed by a 1D-CNN to generate fixed-dimensional feature vectors. (4) Feature Fusion and Classification: The feature vectors from both branches are fused within a neural network, followed by a FC layer for binary classification.

inputs. Through global pooling layers, the network encodes multi-scale information into a fixed-dimensional feature vector, preserving diagnostically relevant signal patterns. These fine-grained temporal features, which may be overlooked in traditional spectrograms, enrich the model’s ability to detect subtle ASD-related indicators.

2.3. Loss Function

To train our multi-branch architecture end-to-end, we employ a binary cross-entropy (BCE) loss, a standard choice for two-class classification tasks. Let $\{(x_i, y_i)\}_{i=1}^N$ be the set of training samples, where x_i denotes the extracted features, and $y_i \in \{0, 1\}$ is the ground-truth label, with 1 indicating ASD and 0 indicating wild-type. Let \hat{p}_i be the predicted probability that sample x_i belongs to the ASD class. The BCE loss is then defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]. \quad (4)$$

By minimising \mathcal{L}_{BCE} , the model refines its decision boundary, thereby improving the probabilistic separation between ASD and WT vocalisations. The BCE loss \mathcal{L}_{BCE} is a widely adopted objective function in binary classification, valued for its computational efficiency and probabilistic interpretability. In our case, \hat{p}_i is computed by the final fully connected layer followed by a sigmoid activation, encapsulating the fused features from the EMD and Spectrogram branches.

3. Dataset and Experimental Setup

3.1. Dataset

The dataset employed in this study was provided by the Interspeech 2025 1st Mice Autism Detection via Ultrasound Vo-

calisation (MADUV) Challenge [26]. It comprises recordings from 84 mice, evenly divided between WT and ASD model groups, with equal representation of male and female subjects. Each mouse underwent a single 5-minute recording session on postnatal day 8 (P8) using an ultrasonic microphone with a sampling rate of 300 kHz. The dataset contains approximately seven hours of audio, capturing intense ultrasonic vocalisations elicited by brief maternal separation.

For robust training and standardised evaluation, the dataset was stratified into training, validation, and test sets, preserving the proportional distribution of WT and ASD mice across genders. Specifically, 60% of the data was allocated for training, 20% for validation, and 20% for testing.

3.2. Experimental Setup

For training, we utilise mini-batch training with a batch size of 8 and the Adam optimiser, setting the learning rate to 10^{-4} and weight decay to 10^{-5} . During the forward pass, each audio is segmented into clips and resampled to 16 kHz. Meanwhile, set the maximum siftings of 200 during the EMD decomposition process to limit the upper iteration, thus generating more stable IMFs. After extracting the EMD and spectral features, the network fuses the two feature vectors through an FC layer and outputs the segment classification result. Ultimately, Subject-level predictions are derived via a majority voting mechanism applied across five clips. Training is conducted for up to 200 epochs, with early stopping activated after 20 epochs without improvement. The optimal model is selected based on the highest subject-level UAR achieved on the validation set. Implemented in PyTorch, we store intermediate EMD and spectrogram features in NumPy .npz files for efficient caching. All experiments were performed on an NVIDIA RTX 4090 GPU.

4. Results and Analysis

4.1. Overall Performance

Table 1 presents the Unweighted Average Recall (UAR) for both segment and subject level classification across various models. The baseline model achieves UAR values of 0.600 and 0.625 for segment and subject level classification, respectively, while our proposed EMD-Spectrogram Fusion Network (ESFN) exhibits superior performance. Notably, ESFN_IMF5, which decomposes signals into five IMFs, attains a subject-level UAR of 0.750, significantly outperforming the baseline, reflecting its distinct learning and understanding of the intrinsic patterns in mice USVs. McNemar’s test [27] was conducted to assess differences between models by examining discordant pairs, where one model classifies correctly while the other does not. A significant imbalance in these pairs indicates a statistically significant performance difference. As has been noted, all p-values shown in Table 1 are below 0.05, confirming that the improvements of the proposed model over the baseline are statistically significant rather than incidental.

4.2. Effect of Different IMF Configurations

Each IMF in the EMD branch represents a primary oscillatory frequency band, and the number of IMFs significantly affects classification performance. While increasing the IMF count broadens frequency coverage, it may also introduce noise and redundancy, undermining stability. Our experiments reveal that ESFN_IMF3, with too few IMFs, fails to capture the multiscale nature of ultrasonic signals (UAR: 0.556 segment-level, 0.750 subject-level), whereas ESFN_IMF8, with excessive IMFs, extracts finer details but amplifies irrelevant components (UAR: 0.631 segment-level, 0.525 subject-level). In contrast, ESFN_IMF5 strikes the optimal balance, achieving UARs of 0.606 (segment-level) and 0.750 (subject-level).

An occlusion heatmap shown in Figure 2 quantifies the contribution of each IMF by measuring classification changes when an IMF is removed. For WT samples, occluding IMF1 or IMF2 increases the likelihood of ASD, suggesting these IMFs encode key WT-specific cues; for ASD samples, IMF1 and IMF4 are most influential, implying a role in amplifying ASD patterns. These findings confirm that IMF contributions vary with segment-specific acoustic properties and, when combined with spectrogram-based global features, enhance the model’s capacity to detect subtle ASD indicators.

4.3. Confusion Matrix Analysis

Table 2 presents the confusion matrix on the validation set, illustrating the model’s capacity to distinguish ASD and WT mice. While segment-level classification exhibits moderate performance, the majority of predictions are correctly distributed, leading to a strong subject-level classification ability exceeding 75%, which shows room for further refinement. Nevertheless, this demonstrates that aggregating segment-level predictions effectively enhances robustness, reinforcing the model’s reliability in ASD detection.

5. Conclusion

This study presented a multi-branch fusion framework that integrates EMD-derived multi-scale oscillatory features with spectrogram-based global time-frequency representations for WT and ASD classification in mice USVs. By addressing the non-stationary and complex nature of USVs, our approach

Table 1: UARs for segment-level and subject-level classification across different models. p-values from McNemar’s test confirm that improvements over the baseline are statistically significant.

Model	Segment-level	Subject-level	p-value
Baseline	.600 (.588 ± .016)	.625 (.588 ± .034)	1.000
ESFN_IMF3	.556 (.575 ± .050)	.750 (.681 ± .013)	0.025
ESFN_IMF8	.631 (.617 ± .025)	.625 (.625 ± .013)	0.034
ESFN_IMF5	.606 (.615 ± .037)	.750 (.713 ± .062)	0.004

Table 2: Confusion matrices (normalised: in [%]) of the two considered analysis levels on the Validation Set. All instances were obtained via ESFN_IMF5 model.

(a) Segment-Level			(b) Subject-Level		
Pred →	WT	ASD	Pred →	WT	ASD
WT	62.2	37.8	WT	77.8	22.2
ASD	35.0	65.0	ASD	25.0	75.0

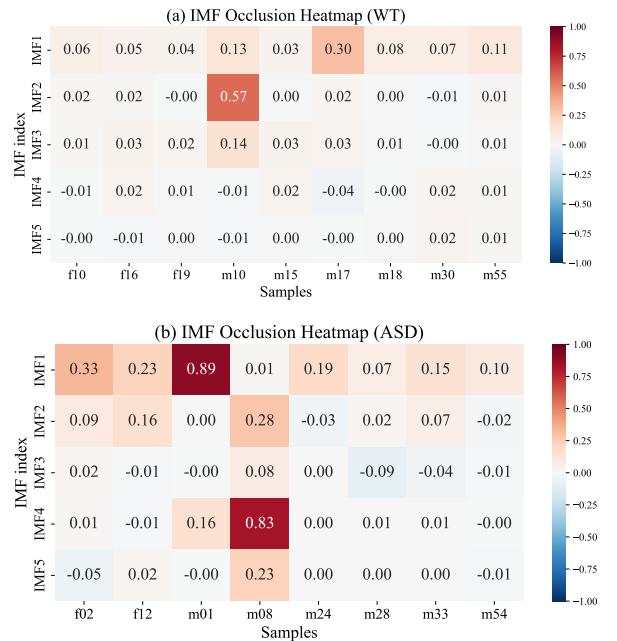


Figure 2: IMF Occlusion Heatmap for WT and ASD Mice. Each cell shows the change in model output when an IMF is occluded. Red values denote an increased likelihood of ASD; blue values, a decreased likelihood. Analysis reveals that IMF1 and IMF2 are key for WT detection, whereas IMF1 and IMF4 are crucial for ASD detection.

outperforms the baseline at both segment and subject levels, demonstrating the complementary strengths of EMD and spectrogram analysis. Beyond performance gains, IMF occlusion analysis reveals that different IMFs contribute uniquely to classification, providing deeper insight into the role of multi-scale features. This underscores the adaptive advantages of EMD in capturing fine-grained vocalisation patterns associated with ASD. These findings validate the robustness and generalisability of our framework, paving the way for further advancements in bioacoustic signal processing. Future research might explore more adaptive decomposition techniques and advanced feature fusion strategies, thereby extending this methodology beyond ASD detection to broader applications in neurological disorder diagnosis and animal communication analysis.

6. Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant Nos.62272044 and 72102068), the National Key R&D Program of China (No.2023YFC2506804), the Beijing Natural Science Foundation (No.L243034), the Ministry of Science and Technology of the People's Republic of China with the STI2030-Major Projects (No.2021ZD0201900), the Japan Society for the Promotion of Science (No.S24116), and the Teli Young Fellow Program from the Beijing Institute of Technology, China.

7. References

- [1] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *The Lancet*, vol. 392, no. 10146, pp. 508–520, 2018.
- [2] Y.-J. Chung, J. Jonkers, H. Kitson, H. Fiegler, S. Humphray, C. Scott, S. Hunt, Y. Yu, I. Nishijima, A. Velds *et al.*, "A whole-genome mouse bac microarray with 1-mb resolution for analysis of dna copy number changes by array comparative genomic hybridization," *Genome Research*, vol. 14, no. 1, pp. 188–196, 2004.
- [3] F. Tian, H. Zhang, Y. Tan, L. Zhu, L. Shen, K. Qian, B. Hu, B. W. Schuller, and Y. Yamamoto, "An on-board executable multi-feature transfer-enhanced fusion model for three-lead eeg sensor-assisted depression diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 1, pp. 152–165, 2025.
- [4] L. Shen, H. Zhang, C. Zhu, R. Li, K. Qian, W. Meng, F. Tian, B. Hu, B. W. Schuller, and Y. Yamamoto, "A first look at generative artificial intelligence based music therapy for mental disorders," *IEEE Transactions on Consumer Electronics*, pp. 1–15, 2024.
- [5] L. Shen, H. Zhang, C. Zhu, R. Li, K. Qian, F. Tian, B. Hu, B. W. Schuller, and Y. Yamamoto, "Enhancing emotion regulation in mental disorder treatment: An aigc-based closed-loop music intervention system," *IEEE Transactions on Affective Computing*, pp. 1–16, 2025.
- [6] B. W. Schuller, J. Löchner, K. Qian, and B. Hu, "Digital mental health—breaking a lance for prevention," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 6, pp. 1584–1588, 2022.
- [7] J. Caston, E. Yon, D. Mellier, H. P. Godfrey, N. Delhaye-Bouchaud, and J. Mariani, "An animal model of autism: behavioural studies in the gs guinea-pig," *European Journal of Neuroscience*, vol. 10, no. 8, pp. 2677–2684, 1998.
- [8] B. Hu, K. Qian, Q. Dong, Y. Luo, Y. Yamamoto, and B. W. Schuller, "Psychological field versus physiological field: From qualitative analysis to quantitative modeling of the mental status," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 5, pp. 1275–1281, 2022.
- [9] M. K. Belmonte, E. Cook, G. M. Anderson, J. L. Rubenstein, W. T. Greenough, A. Beckel-Mitchener, E. Courchesne, L. M. Boulanger, S. B. Powell, P. R. Levitt *et al.*, "Autism as a disorder of neural information processing: directions for research and targets for therapy," *Molecular Psychiatry*, vol. 9, no. 7, pp. 646–663, 2004.
- [10] A. L. Beaudet, "Autism: highly heritable but not inherited," *Nature Medicine*, vol. 13, no. 5, pp. 534–536, 2007.
- [11] J. Nakatani, K. Tamada, F. Hatanaka, S. Ise, H. Ohta, K. Inoue, S. Tomonaga, Y. Watanabe, Y. J. Chung, R. Banerjee *et al.*, "Abnormal behavior in a chromosome-engineered mouse model for human 15q11-13 duplication seen in autism," *Cell*, vol. 137, no. 7, pp. 1235–1246, 2009.
- [12] S. P. Patel, K. Nayar, G. E. Martin, K. Franich, S. Crawford, J. J. Diehl, and M. Losh, "An acoustic characterization of prosodic differences in autism spectrum disorder and first-degree relatives," *Journal of Autism and Developmental Disorders*, vol. 50, pp. 3032–3045, 2020.
- [13] S. Wehrle, F. Cangemi, K. Vogeley, and M. Grice, "New evidence for melodic speech in autism spectrum disorder," in *Proc. Speech Prosody*, vol. 2022, Lisbona, Portugal, 2022, pp. 37–41.
- [14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.
- [15] P. Gao, H. Zhang, L. Shen, Y. Zhang, J. Liu, K. Qian, B. Hu, B. W. Schuller, and Y. Yamamoto, "Clearer lub-dub: A novel approach in heart sound denoising based on transfer learning," in *Proc. Healthcom*. Nara, Japan: IEEE, 2024, pp. 1–6.
- [16] H. Zhang, F. Tian, Y. Tan, L. Shen, J. Liu, J. Liu, K. Qian, Y. Han, G. Su, B. Hu, B. W. Schuller, and Y. Yamamoto, "An ai-assisted all-in-one integrated coronary artery disease diagnosis system using a portable heart sound sensor with an on-board executable lightweight model," *IEEE Transactions on Mobile Computing*, pp. 1–15, 2025.
- [17] E. Marchi, B. Schuller, S. Baron-Cohen, A. Lassalle, H. O'Reilly, D. Pigat, O. Golan, S. Friedenson, S. Tal, S. Bolte *et al.*, "Voice emotion games: Language and emotion in the voice of children with autism spectrum conditio," in *Proc. IDGEL*, New York, USA, 2015, pp. 1–9.
- [18] N. A. Chi, P. Washington, A. Kline, A. Husic, C. Hou, C. He, K. Dunlap, and D. P. Wall, "Classifying autism from crowd-sourced semistructured speech recordings: machine learning model comparison study," *JMIR Pediatrics and Parenting*, vol. 5, no. 2, p. e35406, 2022.
- [19] K. Yao, M. Bergamasco, M. L. Scattoni, and A. P. Vogel, "A review of ultrasonic vocalizations in mice and how they relate to human speech," *The Journal of the Acoustical Society of America*, vol. 154, no. 2, pp. 650–660, 2023.
- [20] A. Ivanenko, P. Watkins, M. A. van Gerven, K. Hammerschmidt, and B. Englitz, "Classifying sex and strain from mouse ultrasonic vocalizations using deep learning," *PLoS Computational Biology*, vol. 16, no. 6, p. e1007918, 2020.
- [21] A. P. Vogel, A. Tsanas, and M. L. Scattoni, "Quantifying ultrasonic mouse vocalizations using acoustic analysis in a supervised statistical machine learning framework," *Scientific Reports*, vol. 9, no. 1, p. 8100, 2019.
- [22] K. Qian, T. Koike, K. Tamada, T. Takumi, B. W. Schuller, and Y. Yamamoto, "Sensing the sounds of silence: A pilot study on the detection of model mice of autism spectrum disorder from ultrasonic vocalisations," in *Proc. EMBC*. Guadalajara, Mexico: IEEE, 2021, pp. 68–71.
- [23] X. Cai and D. Li, "M-edem: A mnn-based empirical decomposition ensemble method for improved time series forecasting," *Knowledge-Based Systems*, vol. 283, p. 111157, 2024.
- [24] X. Cai, D. Li, J. Zhang, and Z. Wu, "Ma-emd: Aligned empirical decomposition for multivariate time-series forecasting," *Expert Systems with Applications*, vol. 267, p. 126080, 2025.
- [25] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.
- [26] Z. Yang, M. Song, X. Jing, H. Zhang, K. Qian, B. Hu, K. Tamada, T. Takumi, B. W. Schuller, and Y. Yamamoto, "Maduv: The 1st interspeech mice autism detection via ultrasound vocalization challenge," 2025. [Online]. Available: <https://arxiv.org/abs/2501.04292>
- [27] M. Q. Pembury Smith and G. D. Ruxton, "Effective use of the mcnemar test," *Behavioral Ecology and Sociobiology*, vol. 74, pp. 1–9, 2020.