



Significance of Time-Frequency preprocessing for automatic Ultrasonic Vocalization classification in Autism Spectrum Disorder model detection

Szymon Szmajdziński¹, Juliusz Wójtowicz-Kruk¹, Ivan Ryzhankow¹, Łukasz Łazarski¹, Jakub Żak¹,
Władysław Średniawa¹

¹Samsung R&D Institute Poland,

w.sredniawa@samsung.com

Abstract

Autism spectrum disorder (ASD) is a complex neurodevelopmental condition of unclear cause and varying severity, often studied using mice as animal models. To accurately distinguish between wild type and ASD model phenotypes, we present a deep learning approach that uses ultrasonic vocalization as input. The proposed method combines a simple model architecture, of convolutional and fully connected layers, with a high-resolution representation of auditory and ultrasound time-frequency patterns. Our approach surpasses baseline performance, achieving an unweighted average recall score of 0.806 on 30-second ultrasonic vocalization fragments. This work was conducted for the 1st INTERSPEECH Mice Autism Detection via Ultrasound Vocalization (MAD-UV) Challenge, where it achieved the highest score among all submitted solutions.

Index Terms: sound recognition, behavioral neuroscience, autism spectrum disorder, ultrasonic vocalization

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by deficits in social communication and repetitive behaviors. While rodent models play a crucial role in studying ASD in humans, confirming the presence of ASD-like phenotypes in genetically modified mice remains challenging due to the complexity and variability of behavioral assays [1, 2]. One aspect of mice behavior is their vocalizations, which occur in both low (below 20 kHz, audible sounds) and high-frequency (about 200 kHz) bands, which are commonly referred to as ultrasonic vocalizations (USV) [3]. USV analysis stands as a promising alternative for detecting ASD-related phenotypes in rodents, as vocal communication being highly relevant to social behavior [4]. Aberrations in USV production have been observed in several ASD mouse models, suggesting a potential link to communication deficits seen in ASD [5, 6, 7]. However, manual or traditional computational methods for USV analysis are time-intensive and prone to human bias, limiting their utility in high-throughput ASD model screening.

Automatic analysis of USVs in mice has gained significant attention, particularly in the context of machine learning applications for vocalization classification [8, 9, 10, 11]. Deep understanding and categorization of these vocalizations are crucial for untangling communication patterns and behavioral states in rodents [12]. Various computational approaches, including machine learning methods, have been employed to classify USVs. Notably, these methods have demonstrated the ability to accurately classify milliseconds-long vocalizations, with accuracies of up to 85% [13]. Recent studies have explored the potential of neural networks to classify sex and strain based on USV features [14]. Interestingly, high prediction accuracy was achieved

by analyzing spectrograms of time windows shorter than 10 ms, highlighting the fine-grained temporal structure of these vocalizations as a key factor in classification performance [14].

In this study, we propose a machine learning-based approach for detecting ASD-like phenotypes in mice based on their USV time-frequency patterns. Our model integrates feature extraction with deep learning classification. We present how we achieved the best model performance on the 1st INTERSPEECH Mice Autism Detection via Ultrasound Vocalization (MAD-UV) Challenge dataset. We show both the best solution from the challenge test set, together with set of alternative solutions and preprocessing methods that also achieved remarkable high scores at our test set. The proposed framework not only improves accuracy in identifying ASD mouse models from wild-type (WT) but also contributes to a deeper understanding of ASD-related vocal communication deficits.

2. Dataset

The dataset provided by the MAD-UV Challenge organizers consists of recordings from 84 mice. These recordings were collected from 44 WT mice (30 males, 14 females) and 40 ASD model mice (27 males, 13 females) [15].

USV recordings were obtained from all mice at postnatal day 8, with each session lasting 5 minutes. High-precision microphones (Avisoft UltraSoundGate 416H) were used for recording at a sampling rate of 300 kHz [15]. The test set consisted of 8 WT (6 males, 2 female) and 8 ASD (6 males, 2 females) subjects. Each test sample was segmented into non-overlapping 30-second clips, resulting in a total of 160 audio samples.

For baseline system implementation, the remaining subjects were further divided into training and validation sets, with 51 subjects allocated to the training set and 17 to the validation set. Detailed information how train and validation set was split for challenge trainings are presented in **Table 1**.

3. Experimental setup

In our experiments, we utilized a consistent pipeline across all models and preprocessing approaches evaluated. We trained and validated the majority of models using the default train/validation split proposed by the challenge organizers, with identical training parameters.

In addition to presenting our best-performing solution, we also report models that achieved strong performance on the validation set but underperformed on the test set. We consider this information valuable for researchers interested in USV classification, highlighting potential challenges and pitfalls that we have already encountered and tested.

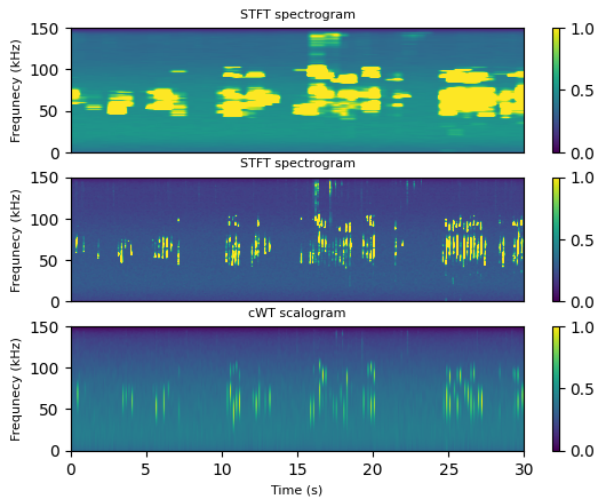


Figure 1: Spectrograms (top and middle) and the wavelet scalogram (bottom) for the same 30-second USV recording segment. The top spectrogram was computed using a 1-second window, while the middle spectrogram was generated with a 100 ms window. The high-resolution spectrogram clearly reveals distinct high-frequency vocalization patterns, which appear blurred when a larger time window is used. Wavelet scalogram preserve fine temporal details of the vocalization patterns. Colorbars encodes normalized power.

3.1. Preprocessing methods

The original 5-minute recordings were segmented into 30-second overlapping clips with a 15-second overlap. We converted each segment into a spectrogram using the Short-Time Fourier Transform (STFT) via `spectrogram` function (0.25 Tukey window) from `Scipy` Python package, version 1.10.1. To ensure consistency in array size regardless of the STFT time window, we subsequently averaged spectrograms across frequencies. In the final step, the logarithm of one plus the averaged spectrogram array was computed, followed by min-max normalization across all frequencies. This pipeline aligns with the baseline method described in [15].

Motivated by findings that higher time-resolution spectrograms improve classification accuracy [13, 14] and by prior research suggesting the potential of wavelet analysis for mice vocalization studies [16], we conducted a wavelet analysis on the segmented waveforms. To efficiently compute the continuous wavelet transform (cWT) across the full frequency spectrum up to 150 kHz, we adapted the implementation from <https://github.com/fastlib/fcwt> [17]. We used Morlet wavelet with a sigma parameter of two. The resulting wavelet scalogram was computed for 500 linearly spaced frequencies between 1 Hz to 150 kHz to ensure compatibility with the baseline spectrogram representations. For final training preparation, the wavelet spectrum in the time domain was smoothed using a 100-sample moving average window and then resampled to reduce scalogram sampling rate from 300 kHz to ~ 100 Hz. This step we found necessary to reduce original size of the scalogram "picture" from 500×9000000 to 500×3000 and thus accommodate CUDA memory constraints for batch. Example spectrograms and scalograms for the same sample from test set are presented in **Figure 1**.

3.2. Architecture and training

We conducted experiments with several model architectures, as outlined below:

1. **Basic Convolutional Neural Network (2xConv+FC):** This model consisted of two convolutional layers followed by a fully connected layer. The filters of the first two convolutional layers were fixed at 32 and 64, respectively. The kernel sizes for the first two convolutional layers were set to (5,5) and (3,3).
2. **Etended CNN with Additional Convolutional Layer (3/4xConv+2xFC):** We added a third convolutional layer with 128 channels, before the fully connected layer of the 2xConv+FC mode. The extra convolutional layer used a kernel size of (3,3).
3. **Transformer-based Model with Swin Transformer and CNN (Swin+Conv):** Inspired by previous work on human ASD speech classification [18], we tested a model based on the Swin transformer combined with a convolutional layer. The Swin transformer is capable of extracting complex, high-level features and attention maps from audio samples. These features are then processed by a convolutional neural network (CNN), which treats the output as two-dimensional input.
4. **Parallel Convolutional Streams with Attention (2x2xConv+Att):** We developed a network architecture consisting of two parallel streams of two convolutional layers, each taking different time-frequency representations as input — one stream for the spectrogram and the other for the cWT scalogram. The outputs of the two streams are concatenated before the fully connected layer, and the combined embeddings are projected through a single attention head. This solution aims to capture the most distinct features from the alternative time-frequency representations.

We evaluated each model to determine its performance in USV classification, assessing the efficacy of different architectural choices. All experiments were run on NVIDIA A100 graphic card, batch size of 48 samples and iterated for 4000 epochs for every training. No model finetuning method was used. For training we used the `PyTorch` (version 1.7.1) framework with the Adam optimizer, loss function: Binary Cross-Entropy (BCE), weight decay parameter of 10^{-6} and a learning rate of 2×10^{-5} . Based on our experiments ResNet-like models quickly overfitted, due to the small dataset and few unique subjects. The challenge's time limited testing more architectures and in consequence we focus on small convolutional networks, working on high resolution spectrograms.

3.3. Metaparameter optimization

The decision threshold is a crucial hyperparameter for binary classification, determining the probability value above which a data point is classified as 1 and below which it is classified as 0. During model training, at each validation epoch, the pipeline calculates the optimal segment-level Unweighted Average Recall (UAR) by performing a grid search over possible threshold values with a step size of 0.05. We select the final model checkpoint based on the highest segment UAR on the validation set, and save it along with the corresponding threshold.

To further validate the optimal threshold for each model, we employed a 4-fold cross-validation approach. Each fold ensures a consistent training-to-validation subject split ratio (train:51, valid:17), maintaining similar distributions of ASD vs. WT and male vs. female subjects. The specific fold assignments are detailed in **Table 1**.

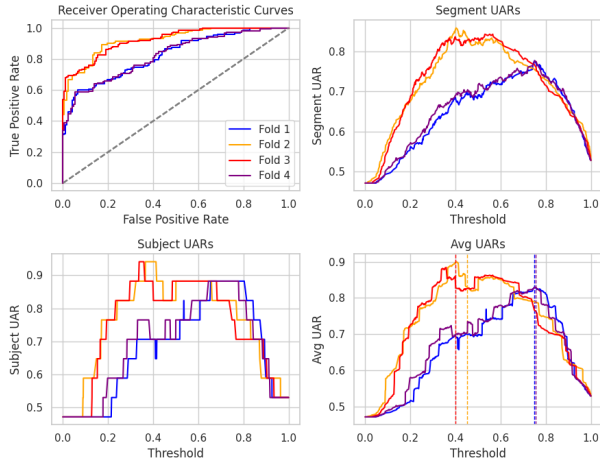


Figure 2: The performance metrics for all folds of the cross-validation are presented as follows. **The upper-left** panel shows the ROC curves, illustrating the trade-off between the False Positive Rate and the True Positive Rate across different threshold values. **The upper-right** panel plots the segment-level Unweighted Average Recalls (UARs) as a function of the threshold. **The lower-left** panel depicts the subject-level UARs against the threshold. Finally, the **lower-right** panel presents the average segment and subject UARs, alongside a comparison to the thresholds determined during training. Notably, folds 1 and 4 achieve their maximum Avg UAR at the predetermined threshold, whereas folds 2 and 3 reach their peak Avg UAR at thresholds lower than those established during the training phase.

Subject-level UAR metric, indicates that the optimal threshold can be lower than the one determined during training. This was due to simplified approach with a step size of 0.05 in the validation. Grid search after training refined the threshold, giving 0.41 and 0.36 for validation fold, with average UAR of 0.90 and 0.88, which outperformed the initial threshold values 0.40 and 0.45 - red and orange dotted lines in **Figure 2**.

Table 1: Cross-validation folds (Total ASD: 32, Males: 45)

ID	Train ASD	Val. ASD	Train males	Val. males
0	24	8	33	12
1	27	5	35	10
2	21	11	34	11
3	24	8	33	12
Total	32		45	

4. Results

4.1. Validation set results

The validation set comprised 17 mice (12 male and 5 female), including 8 ASD subjects, resulting in a total of 318 samples after 30-second segmentation.

The primary evaluation metric was segment-level UAR, with subject-level UAR serving as a secondary measure. According to the challenge evaluation rules, subject-level UAR was strictly derived from segment-level predictions and was computed using majority voting across segments. The high-

Table 2: Results of different models on validation set. **Bolded model is a baseline replication study on validation set.** FT stands for Fourier transform and cWT for wavelet scalogram preprocessing method.

Architecture	Window	Method	Seg. UAR	Sub. UAR
2xConv+FC	0.1 s	FT	0.77	0.88
3xConv+2xFC	0.1 s	FT	0.74	0.94
3xConv+FC	1 s	FT	0.74	0.88
2x2xConv+Att	0.1 s	FT/cWT	0.73	0.89
3xConv+2xFC	1 s	FT	0.69	0.82
4xConv+2xFC	1 s	FT	0.67	0.94
2xConv+FC	1 s	FT	0.67	0.81
Swin+Conv	1 s	FT	0.62	0.72

est segment UAR achieved on the validation set during training was used to identify and save the best model checkpoint.

We present the results for top-performing models in **Table 2**. We found the model with one extra convolutional layer (Conv) and one extra fully connected (FC) layer reached top performance. Larger models (4xConv+2xFC) were overfitting faster on segment level but kept the subject score at 0.94 UAR. Baseline method architecture performed much better on 100 ms window. Based on validation set prediction, we have chosen best models to be tested on challenge test set.

4.2. Challenge test set results

The test set consisted of 160 unlabeled, non-overlapping .wav segments of 8 WT and 8 ASD model mice. According to the evaluation protocol set by the challenge organizers, the best prediction from five submissions was selected, and the final ranking was determined based on the average of segment and subject-level UAR.

To optimize performance under these conditions, we explored various preprocessing and modeling strategies. We found that our top-performing model was a 2xConv+FC architecture trained on 100 ms windowed spectrograms with 50 ms overlap. This model achieved:

- segment-level UAR of **0.806**
- subject-level UAR of **1.0**,

meaning that all mice were correctly classified based on majority voting across their segments. These results demonstrate the effectiveness of a simple model approach and highlight the impact of carefully chosen preprocessing parameters on classification accuracy. It is worth noting that more complex architectures or those with additional convolutional layers performed worse on the test set.

Since the ground truth labels were not disclosed by the organizers, we are unable to report exact performance metrics for our remaining models. However, we observed that most of our models using high-resolution spectrograms outperformed the baseline results of 0.60 for segment-level UAR and 0.625 for subject-level UAR.

4.3. Surrogate test set results

We split our original dataset to create a surrogate test set with known labels. To achieve this, we randomly selected 12 subjects from the training set and 4 subjects from the validation set, ensuring that the original ASD-to-WT proportion and the Female-to-Male ratio were maintained. To better estimate the

overall model performance on the surrogate test set, we computed the average of the three best model checkpoints from validation and grid-searched the best threshold for each checkpoint on the test set.

First, we replicated the training that gave us the best results on the challenge test set, specifically using the 2xConv+FC model on 100 ms windowed segments. For comparison, we also tested the baseline model of the same architecture, but with 1-second windowed spectrograms. Changing the spectrogram window resulted in better average scores for the segment (0.7) and subject UAR (0.75, computed using segment majority voting), compared to 0.67 and 0.7 obtained from the baseline model. Interestingly, further increasing the temporal resolution to 20 and 10 ms led to even better scores:

- **20 ms: 0.81 for segment and 0.92 for subject**
- **10 ms: 0.83 for segment and 0.88 for subject**

Decreasing the window size came at the cost of frequency resolution, which was reduced from 500 to 100 bins. We named spectrograms calculated with 100 bins: FT/5 and 10 bins: FT/50.

Further reducing the frequency resolution from 500 to just 10 bins (model 2xConv+FC, FT/50), while maintaining the temporal resolution of 10 ms, sustained high UAR scores: 0.78 segment and 0.92 subject. It is important to note that extending the number of convolutional layers did not improve scores further. The model achieved 0.69 on the segment level and 0.83 on the subject level (with one additional subject correctly classified). Similarly, using the cWT spectral representation did not lead to improvements, giving scores of 0.64 for the segment level and 0.75 for the subject level. All mean scores for the surrogate test set are presented in **Table 3**.

We also observed that training with high-resolution spectrograms did not lead to rapid overfitting when using convolutional architectures with more than two layers or other larger models. On the other hand, for smaller models, optimal weights were reached after approximately 3,000 training epochs. In general, these findings provide insights into the trade-offs between spectrogram resolution, model complexity, and training dynamics in USV classification.

Table 3: Average scores for different models computed from best three (by validation) training epochs. Bolded model is baseline solution from challenge organizers. FT stands for Fourier transform and cWT for wavelet scalogram preprocessing method. FT/5 and FT/50 are spectrograms with 100 and 10 bins instead of default 500.

Architecture	Window	Method	Seg. UAR	Sub. UAR
2xConv+FC	0.01 s	FT/5	0.83	0.88
2xConv+FC	0.02 s	FT/5	0.81	0.92
2xConv+FC	0.01 s	FT/50	0.78	0.92
3xConv+2xFC	0.02 s	FT/5	0.73	0.83
2xConv+FC	0.1 s	FT	0.7	0.75
3xConv+2xFC	0.1 s	FT	0.69	0.83
2xConv+FC	1 s	FT	0.67	0.7
3xConv+2xFC	0.01 s	cWT	0.64	0.75

5. Discussion

As participants in the 1st INTERSPEECH Mice Autism Detection via Ultrasound Vocalization (MAD-UV) Challenge, we

conducted multiple experiments to optimize binary classification models. Our results demonstrate that Fourier transform spectrograms computed with 100 ms and 20 ms windows outperformed those using a 1-second window on both the challenge and surrogate test sets. This finding is consistent with previous research showing that key features of rodent vocalizations occur on a millisecond scale [19]. It has also been estimated that mice and rats emit USVs with a mean call duration of 80 ms in various social contexts, which explains why shorter time windows are more suitable for distinguishing nuances in ASD mice model vocalizations [20].

We hypothesized that further improvements could be achieved by reducing the window size and increasing the complexity of the model. However, previous studies on multi-class USV classification have primarily used models with few convolutional and fully connected layers, similar to our top-performing architecture [21]. Our experiments, conducted with an extended number of convolutional layers, confirm that a larger model does not necessarily achieve better predictions, especially as suggested for small datasets [22]. Investigation of the performance of the model on the surrogate test set has shown that the temporal distribution of USV patterns is a crucial classification factor. Reducing the number of frequency bins to 10 while maintaining the 10 ms spectral windows still allows the model to perform at high classification level, showing that temporal resolution is a key to distinguish between the WT and ASD mice model.

We believe that a proper frequency-time resolution trade-off must be maintained for accurate ASD classification. We further propose that general USV features, such as timing or vocalization length, could be the most relevant for ASD model classification. However, considering previous research by Mohrle et al. [23], which discusses complex relations of the vocalization of mouse pups, it merits further study. Our findings indicate that automated USV analysis can serve as a valuable tool to detect ASD-like traits in genetic mouse models, bridging behavioral neuroscience with computational methodologies.

6. Conclusions and limitations

In our study and challenge work, we assessed classic deep learning to improve the reliability of ASD mice model identification, with potential implications for preclinical research. Our solution achieved first place in the final challenge ranking for both segment and subject level classification. We trained compact models that outperformed the baseline when applied to the full spectrum using 100 ms and shorter spectrogram windows. However, more complex architectures, including those with additional layers or attention mechanisms, did not yield significant improvements.

Furthermore, the adapted method tested for human ASD speech classification did not perform well for mouse vocalizations, suggesting that different signal processing models may be required. Using computationally costly but high-resolution wavelet representations also failed to improve model predictions, even when trained alongside spectrograms. This may be due to the need for a more complex architecture for wavelets or the loss of important USV information caused by resampling operations to fit our hardware.

We believe that the observations and results from our study can aid future model development for both USV patterns and ASD mice model classification.

7. References

- [1] J. L. Silverman, M. Yang, C. Lord, and J. N. Crawley, "Behavioural phenotyping assays for mouse models of autism," *Nature Reviews Neuroscience*, vol. 11, pp. 490–502, 7 2010.
- [2] J. Heckman, B. McGuinness, T. Celikel, and B. Englitz, "Determinants of the mouse ultrasonic vocal structure and repertoire," *Neuroscience and Biobehavioral Reviews*, vol. 65, pp. 313–325, 6 2016.
- [3] M. L. Dent, R. R. Fay, and A. N. Popper, *Rodent bioacoustics*. Springer, 2018, vol. 67.
- [4] M. Wöhr, "Ultrasonic vocalizations in shank mouse models for autism spectrum disorders: Detailed spectrographic analyses and developmental profiles," *Neuroscience and Biobehavioral Reviews*, vol. 43, pp. 199–212, 6 2014.
- [5] E. Fujita, Y. Tanabe, A. Shiota, M. Ueda, K. Suwa, M. Y. Momoi, and T. Momoi, "Ultrasonic vocalization impairment of foxp2 (r552h) knockin mice related to speech-language disorder and abnormality of purkinje cells," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 3117–3122, 2 2008.
- [6] M. Wöhr, F. I. Rouillet, A. Y. Hung, M. Sheng, and J. N. Crawley, "Communication impairments in mice lacking shank1: Reduced levels of ultrasonic vocalizations and scent marking behavior," *PLoS ONE*, vol. 6, p. e20631, 6 2011.
- [7] S. Roy, N. Watkins, and D. Heck, "Comprehensive analysis of ultrasonic vocalizations in a mouse model of fragile x syndrome reveals limited, call type specific deficits," *PLoS ONE*, vol. 7, p. e44816, 9 2012.
- [8] M. Van Segbroeck, A. T. Knoll, P. Levitt, and S. Narayanan, "Muppet—mouse ultrasonic profile extraction: a signal processing tool for rapid and unsupervised analysis of ultrasonic vocalizations," *Neuron*, vol. 94, no. 3, pp. 465–485, 2017.
- [9] K. R. Coffey, R. E. Marx, and J. F. Neumaier, "Deepsqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations," *Neuropsychopharmacology*, vol. 44, pp. 859–868, 4 2019.
- [10] R. O. Tachibana, K. Kanno, S. Okabe, K. I. Kobayashi, and K. Okanoya, "Usvseg: A robust method for segmentation of ultrasonic vocalizations in rodents," *PLoS one*, vol. 15, no. 2, p. e0228907, 2020.
- [11] K. J. Scott, L. J. Speers, and D. K. Bilkey, "Utilizing synthetic training data for the supervised classification of rat ultrasonic vocalizations," *The Journal of the Acoustical Society of America*, vol. 155, pp. 306–314, 1 2024.
- [12] D. T. Sangiamo, M. R. Warren, and J. P. Neunuebel, "Ultrasonic signals associated with different types of social behavior of mice," *Nature Neuroscience*, vol. 23, pp. 411–422, 3 2020.
- [13] 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, p. 8100, 5 2019.
- [14] A. Ivanenko, P. Watkins, M. A. J. van Gerven, K. Hammer-schmidt, and B. Englitz, "Classifying sex and strain from mouse ultrasonic vocalizations using deep learning," *PLOS Computational Biology*, vol. 16, p. e1007918, 6 2020.
- [15] Z. Yang, M. Song, X. Jing, H. Zhang, K. Qian, B. Hu, K. Tamada, T. Takumi, B. W. Schuller, and Y. Yamamoto, "Mad-uv: The 1st interspeech mice autism detection via ultrasound vocalization challenge," *arXiv preprint arXiv:2501.04292*, 2025.
- [16] A. A. Smith and D. Kristensen, "Deep learning to extract laboratory mouse ultrasonic vocalizations from scalograms," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1972–1979.
- [17] L. P. A. Arts and E. L. van den Broek, "The fast continuous wavelet transformation (fcwt) for real-time, high-quality, noise-resistant time–frequency analysis," *Nature Computational Science*, vol. 2, pp. 47–58, 1 2022.
- [18] A. Jaby and M. B. Islam, "Audio speech signal analysis for early autism spectrum disorder detection," in *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2023, pp. 1–6.
- [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, pp. 650–660, 8 2023.
- [20] M. L. Scattoni, C. Michetti, and L. Ricceri, *Rodent Vocalization Studies in Animal Models of the Autism Spectrum Disorder*, 2018, pp. 445–456.
- [21] A. H. Fonseca, G. M. Santana, G. M. Bosque Ortiz, S. Bampi, and M. O. Dietrich, "Analysis of ultrasonic vocalizations from mice using computer vision and machine learning," *eLife*, vol. 10, p. e59161, mar 2021. [Online]. Available: <https://doi.org/10.7554/eLife.59161>
- [22] J. Zhou, Q. He, G. Cheng, and Z. Lin, "Union-net: lightweight deep neural network model suitable for small data sets," *The Journal of Supercomputing*, vol. 79, pp. 7228–7243, 5 2023.
- [23] D. Möhrle, M. Yuen, A. Zheng, F. L. Haddad, B. L. Allman, and S. Schmid, "Characterizing maternal isolation-induced ultrasonic vocalizations in a gene–environment interaction rat model for autism," *Genes, Brain and Behavior*, vol. 22, 6 2023.