

# The NPU-ASLP System for Deepfake Algorithm Recognition in ADD 2023 Challenge

Ziqian Wang, Qing Wang, Jixun Yao and Lei Xie\*

Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xian, China

## Abstract

This paper describes our NPU-ASLP system for the Deepfake Algorithm Recognition (AR) task in the Audio Deepfake Detection 2023 Challenge. This task is an open-set classification problem focusing on identifying the specific algorithms used to create the deepfake speech utterances. In this task, we introduce a deepfake AR system with contributions in data augmentation, model architecture, fine-tuning strategy, and model ensemble. We first generate training data by applying various data augmentation techniques to the deepfake speech. We then utilize ResNet101 and a long-term temporal-frequency transformer module to better capture audio context dependencies. Moreover, we employ pre-trained WavLM for better feature extraction. Additionally, our content-invariant fine-tuning strategy improves performance. Finally, model ensemble with different representation combinations further enhances performance. Experiments show that our system achieves an F1-score of 0.7355 on the evaluation set, and ranks fourth in the challenge.

## Keywords

Deepfake algorithm recognition, data augmentation, transformer, model ensemble

## 1. Introduction

In recent years, the progress of deep learning has resulted in significant advancements in speech synthesis [1] and voice conversion [2] technologies. These technologies are capable of producing speech that is highly realistic and bears a strong resemblance to natural human speech. However, if misused, these technologies have the potential to cause harm to society. Therefore, detecting deepfake audio, which is manipulated audio created by deep learning algorithms, has become an urgent task to prevent any potential misuse.

The objective of the second edition of the Audio Deepfake Detection Challenge (ADD 2023) [3] is to inspire researchers to develop innovative and pioneering technologies that can boost and encourage further exploration in identifying and analyzing deepfake speech utterances. Unlike earlier challenges (such as the previous edition of ADD challenge [4], and the ASVspoof challenge [5, 6, 7]), ADD 2023 aims to transcend the limitations of binary real/fake classification and instead focus on identifying the exact regions that are manipulated in a partially fake speech, as well as identifying the source responsible for generating any fake audio. Additionally, ADD 2023 comprises multiple evaluation rounds for the fake audio game sub-challenge.

The Deepfake Algorithm Recognition (AR) task in ADD 2023 is an essential challenge that focuses on iden-

tifying the specific algorithms used to create deepfake speech utterances. With the rapid advancements in deep learning as speech synthesis and voice conversion, deepfake audio [8, 9] has become an increasingly significant concern. It has the potential to be misused for various malicious purposes, such as spreading false information, creating non-consensual audio, and impersonating individuals. Furthermore, this task is complicated by the fact that deepfake algorithms can be trained on a wide variety of data sources, such as real speech samples, synthesized speech, or a combination of both. Additionally, deepfake audio can be generated using a variety of methods, leveraging the recent advancements in deep learning based audio generation [10, 11, 12]. These complexities make the development of effective deepfake detection algorithms a challenging problem.


This report describes our system for the deepfake AR task. We shape this task as an open set classification problem [13], due to the evaluation set containing an unknown counterfeit. Specifically, we introduce a threshold-based classification and detection system with contributions in four primary areas, including augmenting the available data, enhancing the model architecture, adopting a content-invariant finetuning strategy, and utilizing model ensemble techniques. First, data augmentation approaches are investigated to expand the official dataset. We adopt the noise and room impulse response clips and simulate training data by adding distortions to the deepfake speech utterances. We further apply random sampling, time stretching, time and frequency masking to the mixed audio to better reveal the artifacts in the deepfake speech utterances. As for the model architecture, we employ ResNet101 [14] with a Temporal-Frequency Transformer (TFT) module as the training network based

*IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R*

\*Corresponding author.

✉ zq\_wang@mail.nwpu.edu.cn (Z. Wang)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

on the augmented data, which better captures long-term temporal-frequency dependencies and further aggregates global hierarchical contextual information. Moreover, we leverage the pre-trained audio-LLM WavLM [15] to facilitate versatile latent representations. Additionally, we explore feature concatenation and multi-channel representation fusion as feature combination techniques. Finally, we utilize model ensemble to improve the robustness and generalization ability of our system.

## 2. Deepfake AR System

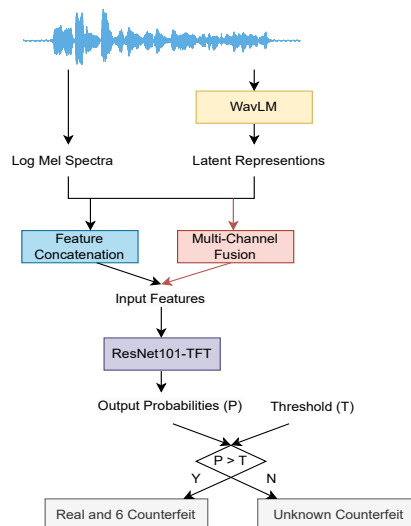
### 2.1. System Overview

Our system comprises three main modules designed to improve the performance of deepfake audio detection. We select ResNet101 [14] as the backbone network due to its exceptional capabilities in various classification tasks. To better capture contextual dependencies, we introduce a transformer [16] based Temporal-frequency modeling module dubbed TFT. In addition, we leverage the pre-trained audio-LLM WavLM [15] to extract latent representations, thereby enhancing the compactness and generalization of the input features.

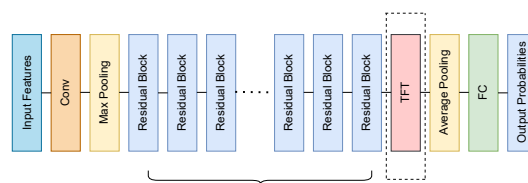
When inputting an audio to be detected, log mel spectra and latent representations are extracted first. Two alternative combination techniques, namely feature concatenation and multi-channel fusion, are employed to derive the input features. The combined features are then fed into the ResNet101-TFT network to generate predicted probabilities. To determine the classification, we establish a hyperparameter threshold denoted as  $T$  and compare the maximum probability of prediction against this threshold. If the maximum probability is below  $T$ , the prediction is labeled as an unknown counterfeit, while exceeding the threshold identifies it as belonging to a specific category. The details are illustrated in Figure 1.

### 2.2. Backbone Network

ResNet [17] is a well-known convolutional neural network (CNN) based network that have achieved great performance in different tasks, the use of residual connections allows information from earlier layers of the network to be easily passed forward to later layers, addressing the problem of vanishing gradients in deep networks. To this end, we utilize ResNet101 as the backbone. The structure of ResNet 101 as shown in Figure 2, is composed of a series of building blocks, each of which consists of multiple convolutional layers and a skip connection that bypasses the convolutional layers. The first block in ResNet101 consists of a single convolutional layer followed by a max pooling layer, while the remaining blocks each contain multiple residual blocks. The final layer of the network is a fully connected layer with



**Figure 1:** The diagram of our deepfake AR system pipeline. Log Mel Spectra and latent representations are incorporated in two alternative combination techniques to obtain input features, which are represented by black directional connector and red directional connector, respectively.

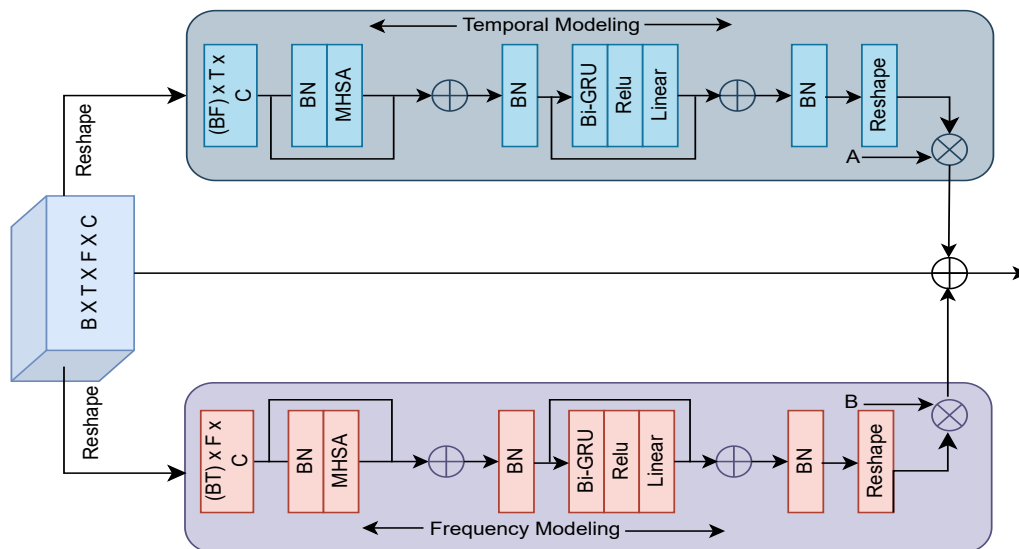


**Figure 2:** The diagram of ResNet101-TFT. The dashed line portion represents the TFT module, while the remaining sections adhere to the original ResNet101 architecture.

softmax activation, which is used to produce the output class probabilities.

### 2.3. TFT Module

Speech signals are inherently long sequences of data that exhibit time-varying and periodic characteristics. To effectively capture long-term temporal dependencies and aggregate contextual information across different frequency bands, inspired by [18, 19], we introduce the Temporal-Frequency Transformer (TFT) module. By incorporating the TFT module, we aim to enhance the ability of our system to capture complex temporal dynamics and exploit the contextual relationships in deepfake speech utterances. TFT includes two branches that enable the modeling of long-term dependencies along both the temporal and frequency dimensions. These branches, called the adaptive temporal modeling branch and the



**Figure 3:** The diagram of the TFT module. The temporal modeling branch and the frequency modeling branch, individually capture temporal and frequency information to effectively facilitate long-term contextual dependencies.

adaptive frequency modeling branch, use two adaptive weights  $A$  and  $B$ . The module incorporates an improved transformer, which includes a multi-head self-attention (MHSA) module and a GRU-based feed-forward network. Residual connections and batch normalization are also employed. To better model long-term dependencies, the feed-forward network uses a bi-directional GRU (Bi-GRU) instead of the first fully connected layer in the traditional transformer [16]. These changes allow the TFT module to capture long-term dependencies more effectively. A diagram of the TFT module is shown in Figure 3.

#### 2.4. Utilization of Pre-trained WavLM

Recently, following the success of the language model BERT on textual data representation, various models which learn audio data representation have been developed. The objective is to train a system for a single task requiring an extensive representation of the underlying audio data. Once the network is trained, the last few layers are then removed to get a system building an extensive vectorial representation of audio data, which can then be used as input features. WavLM [15] is the latest self-supervised model built with transformer blocks trained on Mix94k, a corpus of 94k hours drawn from LibriLight, VoxPopuli and GigaSpeech. WavLM learns to represent speech by masking a part of the signal and trying to predict the hidden part. On this aspect, this model is similar to the self-supervised models HuBERT [20] and wav2vec2.0 [21]. WavLM is task agnostic and achieves state-of-the-art performance on various benchmarks. To

this end, we employ WavLM Base+ as the pre-trained model and freeze its weights to extract high-level general latent representations.

### 3. Experiments

#### 3.1. Datasets

The training and development datasets of ADD 2023 Challenge Task 3 consist of 22,397 and 8,400 audio clips, respectively. We rearrange the training set and development set in a ratio of 9 to 1. All audios are in single-channel 16KHz 16-bit format, ranging from 1 to 105 seconds. In the training and development datasets, there are a total of 7 categories, with 1 category representing real audio and the remaining 6 categories representing counterfeit audio. Each of these categories is labeled from 0 to 6. The test dataset, on the other hand, comprises 8 categories, including the 7 categories found in the training and development datasets and an additional category representing an unknown counterfeit. To better leverage the training data, we clip long-duration deepfake speech utterances into segments ranging from 4 to 8 seconds with the same label. ADD 2023 Challenge permits the use of external non-speech resources (e.g. noise samples and impulse responses), therefore 50,000 noise and room impulse response audio clips are sampled from the 4th DNS challenge [22] and utilized in our training pipeline.

### 3.2. Data Augmentation

To further reveal the artifacts in the deepfake speech utterances and make the deep learning based model robust, our data augmentation pipeline comprises two strategies. One is that we generate mixed audio by mixing speeches, noises, and room impulse responses. One original deepfake speech audio is added by a piece of noise recording with a signal-to-noise ratio (SNR) drawn from a Gaussian distribution with  $N(-5, 20)$ dB with an 80% possibility. Then, this audio is convolved with a piece of room impulse response with a possibility of 50%. The other is that random sampling, time stretching, time, and frequency masking are applied to the mix audios generated by the first strategy with a certain probability. Specifically, one mix audio is upsampled or downsampled ranging in [8kHz, 16kHz, 24kHz], then processed with speed shifting and temporal and frequency masking.

### 3.3. Preprocessing

We extract input features from the augmented audios in different hierarchies. To be exact, we facilitate 128-dimensional log Mel spectrograms extracted with 1024 samples of Hann window and 512 samples shift as low-level common features  $\mathbb{F}_{low}^{T \times F \times 1}$ , we further exploit the pre-trained WavLM to extract high level general latent representations  $\mathbb{R}_{high}^{T \times F \times 1}$ . We explore two different feature combination techniques: one is concatenating low-level features and high-level features along the feature dim as feature concatenation resulting in the input features  $\mathbb{I}^{T \times F \times 1}$ , the other is combining low-level features and high-level features as multi-channel representation fusion resulting in the input features  $\mathbb{I}^{T \times F \times 2}$ , in which T denotes time-frames and F denotes the feature dim.

### 3.4. Training Details

The augmented training data are clipped into 4-sec long segments. AdamW optimizer is adopted to train the model and ReduceLROnPlateau learning rate scheduler is applied to control the learning rate. Batch size is set to 48 and the upper limit of training epochs is set to 100. The optimization criteria are Cross Entropy loss and the Kullback-Leibler divergence loss [23] between the prediction probabilities and target labels as we set the output size of the last fully connected layer in our deepfake AR system to be 7, the highest logit indicates the predicted class. The targets are set to be the same size, where the probability of the true class is 1 and the others are 0. The hyperparameters  $A$  and  $B$  in the TFT module are initialized to 1. Early stop is utilized when the accuracy rate on the dev set stops improving for 10 epochs. gradient clipping is employed to avoid overfitting and encourage the models to learn more robust and

generalizable representations.

### 3.5. Finetuning Strategy

We intuitively perceive the Deepfake Algorithm Recognition as a content-invariant classification task, due to the fact that humans tend to assess the authenticity of an audio through aspects such as pitch, tone, naturalness, emotions, and subjective auditory perception, rather than the actual content of the audio. To this end, we draw on the experience of the mixup strategy [24, 25, 26] from computer vision and introduce a content-invariant finetuning strategy. Specifically, we fully exploit the official dataset provided by ADD Challenge in two ways. One is that we simulate multi-speaker audios by mixing different deepfake utterances with the same label with a signal-to-noise ratio (SNR) drawn from a Gaussian distribution with  $N(-5, 20)$ dB, and the other is that we randomly clip deepfake utterances with the same label into 1-2 secs and concatenate together to obtain new audio sequences. We then employ the generated data to finetune our model and achieve better results on the evaluation set, the details are discussed in Section 4.

### 3.6. Model Ensemble

Model ensemble [27, 28] is a machine learning technique that combines the predictions of multiple models to achieve better performance than any individual model, which reduces the impact of individual model biases and errors and improves the overall accuracy and robustness of the predictions.

As the test dataset includes an additional category for an unknown counterfeit, which is not present in the training and development datasets, we set the output size of our deepfake AR system to 7. To detect the unknown counterfeit, we compare the posterior probability output  $P$  with a hyperparameter threshold  $T$ . If the posterior probability  $P$  is less than the threshold  $T$ , we consider the audio to be the unknown counterfeit, otherwise, the audio is classified into the 7 known categories. The hyperparameter  $T$  is tuned by the performances on the evaluation set.

$$y_{ens} = \sum_{i=1}^n \frac{w_i}{\sum_{j=1}^n w_j + \lambda} y_i + \epsilon \quad (1)$$

In model ensemble, we use weighted averaging [29] to combine the predictions of multiple models and obtain the ensemble posterior probability, demonstrated in Equation (1). The ensemble prediction probability is calculated by taking a weighted average of the individual model predictions. The weights  $w_i$  are assigned based on the accuracy of each model on the development set, which ensures that the more accurate models are given

**Table 1**

Experimental results of the proposed methods for Evaluation Dataset.

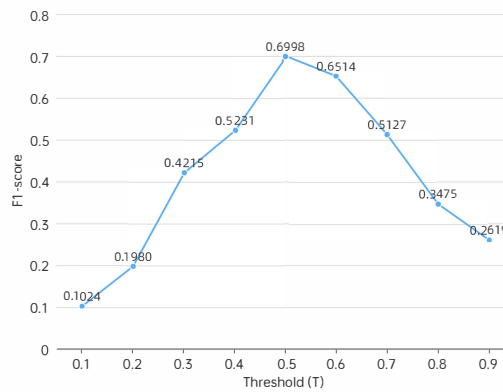
|                     | Threshold (T) | F1-score      |
|---------------------|---------------|---------------|
| ResNet101           | 0.55          | 0.6578        |
| ResNet101-TFT       | 0.53          | 0.6831        |
| ResNet101-TFT-FC    | 0.45          | 0.6996        |
| /w mixup finetuning | 0.45          | 0.7121        |
| ResNet101-TFT-MF    | 0.50          | 0.6998        |
| /w mixup finetuning | 0.50          | 0.7136        |
| Ensemble            | 0.47          | <b>0.7355</b> |

greater importance in the final prediction. we add a small constant  $\epsilon$  to prevent the possibility of a zero-weighted prediction. A regularization parameter  $\lambda$  that controls the strength of the L2 regularization is applied to help to prevent any one weight from dominating the final prediction. This weighting scheme allows us to effectively leverage the strengths of individual models while mitigating the impact of potential biases or errors that may be present in any single model. The use of weighted averaging represents a powerful technique for model ensemble that can improve the accuracy and robustness of our proposed deepfake AR system.

## 4. Result & Discussion

Table 1 presents the experimental results of various deepfake algorithm recognition systems using the F1-score metric, the threshold column in the table denotes the value at which the model achieves the best performance on the evaluation set, the model’s performance declines when the threshold exceeds or falls below this optimal value, as an example is shown in Figure 4. Two ResNet101 variants, ResNet101-TFT-FC and ResNet101-TFT-MF, are exploited with different feature combination techniques: feature concatenation (FC) and multi-channel representation fusion (MF), respectively, using the Temporal-Frequency Transformer (TFT) module.

The results indicate that the inclusion of the Temporal-Frequency Transformer (TFT) module in ResNet101 enhances the system’s capability, and the utilization of feature combination techniques provides additional benefits. Applying the mixup finetune strategy improves the performance of both variants, which aligns with the perception of AR as a content-invariant classification task. The ensemble method, which uses weighted averaging of the outputs of ResNet101-TFT-FC and ResNet101-TFT-MF with mixup finetuning, achieves the best performance, with an F1-score of 0.7355.


**Figure 4:** F1 score of ResNet101-TFT-MF on the evaluation set under different threshold T conditions.

## 5. Conclusion

In this report, we introduce the NPU-ASLP system for the Deepfake Algorithm Recognition task in the Audio Deepfake Detection Challenge. The main contributions of our method are data augmentation, a more powerful model architecture with the TFT module and the application of pre-trained WavLM, content-invariant finetuning strategy, and model ensemble. The experimental results show that the proposed system achieves an F1-score of 0.7355 on the evaluation set and attains fourth place in the challenge.

## References

- [1] X. Tan, T. Qin, F. Soong, T.-Y. Liu, A survey on neural speech synthesis, arXiv preprint arXiv:2106.15561 (2021).
- [2] B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2020) 132–157.
- [3] J. Yi, J. Tao, X. Y. Ruibo Fu, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, H. Li, Add 2023: the second audio deepfake detection challenge, in: *IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)*, 2023.
- [4] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, H. Li, Z. Lian, B. Liu, Add 2022: the first audio deep synthesis detection challenge, 2022. arXiv:2202.08433.

- [5] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, *Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge*, in: Sixteenth annual conference of the international speech communication association, 2015.
- [6] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K. A. Lee, *The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection* (2017).
- [7] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, K. A. Lee, *Asvspoof 2019: Future horizons in spoofed and fake audio detection*, arXiv preprint arXiv:1904.05441 (2019).
- [8] S. Lyu, *Deepfake detection: Current challenges and next steps*, in: 2020 IEEE international conference on multimedia & expo workshops (ICMEW), IEEE, 2020, pp. 1–6.
- [9] Z. Almutairi, H. Elgibreen, *A review of modern audio deepfake detection methods: Challenges and future directions*, *Algorithms* 15 (2022) 155.
- [10] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, *Grad-tts: A diffusion probabilistic model for text-to-speech*, in: International Conference on Machine Learning, PMLR, 2021, pp. 8599–8608.
- [11] J. Kim, S. Kim, J. Kong, S. Yoon, *Glow-tts: A generative flow for text-to-speech via monotonic alignment search*, *Advances in Neural Information Processing Systems* 33 (2020) 8067–8077.
- [12] T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, *CycleGAN-vc2: Improved cycleGAN-based non-parallel voice conversion*, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6820–6824.
- [13] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, T. E. Boulton, *Toward open set recognition*, *IEEE transactions on pattern analysis and machine intelligence* 35 (2012) 1757–1772.
- [14] Q. Zhang, *A novel resnet101 model based on dense dilated convolution for image classification*, *SN Applied Sciences* 4 (2022) 1–13.
- [15] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., *Wavlm: Large-scale self-supervised pre-training for full stack speech processing*, *IEEE Journal of Selected Topics in Signal Processing* 16 (2022) 1505–1518.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, et al., *Attention is all you need*, in: Proc. NeurIPS, 2017, pp. 5998–6008.
- [17] K. He, X. Zhang, S. Ren, J. Sun, *Deep residual learning for image recognition*, in: Proc. CVPR, 2016, pp. 770–778.
- [18] S. Zhang, Z. Wang, J. Sun, Y. Fu, B. Tian, Q. Fu, L. Xie, *Multi-task deep residual echo suppression with echo-aware loss*, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 9127–9131.
- [19] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, H. Wang, *Dual-branch attention-in-attention transformer for single-channel speech enhancement*, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7847–7851.
- [20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021) 3451–3460.
- [21] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, *wav2vec 2.0: A framework for self-supervised learning of speech representations*, *Advances in neural information processing systems* 33 (2020) 12449–12460.
- [22] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusseych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, R. Aichner, *ICASSP 2022 deep noise suppression challenge*, 2022. arXiv:2202.13288.
- [23] J. M. Joyce, *Kullback-leibler divergence*, in: International encyclopedia of statistical science, Springer, 2011, pp. 720–722.
- [24] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, *mixup: Beyond empirical risk minimization*, arXiv preprint arXiv:1710.09412 (2017).
- [25] H. Inoue, *Data augmentation by pairing samples for images classification*, arXiv preprint arXiv:1801.02929 (2018).
- [26] C. Si, Z. Zhang, F. Qi, Z. Liu, Y. Wang, Q. Liu, M. Sun, *Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning*, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 1569–1576.
- [27] Y. Xiao, J. Wu, Z. Lin, X. Zhao, *A deep learning-based multi-model ensemble method for cancer prediction*, *Computer methods and programs in biomedicine* 153 (2018) 1–9.
- [28] M. A. Ganaie, M. Hu, A. Malik, M. Tanveer, P. Suganthan, *Ensemble deep learning: A review*, *Engineering Applications of Artificial Intelligence* 115 (2022) 105151.
- [29] D. Card, M. Zhang, N. A. Smith, *Deep weighted averaging classifiers*, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 369–378.