Exploring Generalization To Unseen Audio Data For Spoofing: Insights From SSL Models

Atharva Kulkarni, Hoan My Tran, Ajinkya Kulkarni, Sandipana Dowerah, Damien Lolive, Mathew Magimai Doss

Exploring Generalization To Unseen Audio Data For Spoofing: Insights From SSL Models

Atharva Kulkarni[*,‡], Hoan My Tran[*,†], Ajinkya Kulkarni[§],
Sandipana Dowerah[†], Damien Lolive[†], Mathew Magimai Doss[§]

IDIAP, Switzerland[§], Univ. Rennes, CNRS, IRISA, France[†], MBZUAI, UAE[‡]

atharva7kulkarni@gmail.com, hoan.tran@irisa.fr, ajinkya.kulkarni@idiap.ch

sandipana.dowerah@irisa.fr, damien.lolive@irisa.fr, mathew.magimaidoss@idiap.ch

Abstract

Deep learning-based speech synthesis has significantly improved realistic audio deepfakes. Despite advanced techniques such as self-supervised learning (SSL) and datasets, current state-of-the-art (SOTA) detection systems fail in out-of-domain scenarios due to the inability to generalize. This work explores the generalization problem through comprehensive experimentation on cross-data evaluation. We observed how training data impacts model generalization, revealing that even SOTA systems struggle with consistent performance across different evaluation settings. This indicates a lack of extensive generalization abilities, especially in SSL approaches. To address this problem, we propose a multi-stage training framework alongside an ensemble of different systems to enhance the robustness and reliable detection in known and unknown out-of-domain scenarios. Experimental evaluation underscores the importance of an ensemble approach to mitigate the limitations in individual systems.

1. Introduction

The advancement in biometric technology has also advanced the generation of deepfakes. 'Deepfakes' refers to creating realistic images, videos, or audio of a person's likeness with another person using deep learning techniques [1]. Among these, the generation of audio deepfakes has become increasingly sophisticated and is becoming hard to spot.

Audio deepfakes fall into two categories, Text-to-speech (TTS) and voice conversion (VC), where the significant difference lies in the input. In the case of TTS, a text is transformed into speech resembling a human voice using technology. On the other hand, voice conversion takes an individual's voice and alters it to sound like another person's while preserving the original speech's linguistic attributes. Deep learning-based TTS and VC systems have made remarkable progress over the years [2, 3]. These technologies generate highly natural-sounding speech that is challenging to differentiate from authentic audio. While these technologies offer various conveniences in our day-to-day lives, such as virtual assistants, translation services, and navigation services, they also seriously threaten social security. With the easy access of deepfake tools [1], voice cloning is becoming easier every passing day. One of the most notable instances of audio deepfakes involved a series of robocalls in New Hampshire that mimicked President Joe Biden's voice [4, 5]. Notably, it took only 20 minutes and 1 US dollar to generate the fake audio of President Biden, discouraging voters to cast their ballots by a street magician [6].

In the digital era, where privacy is becoming increasingly crucial, it is imperative to develop robust detection mechanisms to detect authentic media from doctored media. In recent years, there has been a growing number of efforts aimed at advancing the field of audio deepfake detection [7, 8, 9, 10, 11]. Most of these studies either include a front-end feature extractor and a back-end classifier, which has been the standard framework for many years, or an end-to-end approach utilizing a model that jointly optimizes feature extraction and classification by directly processing raw audio waveforms [12]. However, despite the promising performance of previous studies on audio deepfake detection, the research remains largely fragmented, with few comprehensive surveys. Most studies summarise previous spoofing attacks and countermeasures to protect automatic speaker verification (ASV) systems. Wu et al. [13] provided a comprehensive survey in 2015 assessing the vulnerability of ASV systems and the countermeasures to protect them. The ASVspoof challenges [14] have been crucial in promoting research on detecting spoofed speech to protect ASV systems from manipulation. Kamble et al. [15] discuss advances in anti-spoofing from the perspective of ASVspoof challenges in 2020. Tan et al. [16] analyzes attack detection work for ASV systems published between 2015 and 2021. Mittal et al. [17] review and analyze benchmark spoofed speech datasets, methods, and evaluation metrics for ASV systems and spoof detection techniques.

Although the ASVspoof initiatives appear to demonstrate substantial advancements, with increasingly lower state-of-the-art error rates being frequently reported [18, 19]. The effectiveness of these solutions in real-world situations often remains unproven [20, 21]. There are concerns that these systems struggle to generalize to out-of-domain scenarios. Specifically, they exhibit limited generalization capabilities when faced with deepfakes created using new or different attack algorithms or on unseen data compared to those used in their training data, which has been and continues to be a significant concern [22]. In the ASVspoof 2019 challenge, in the logical access (LA) track, the organizers ensured the assessment of spoofing detection against unknown spoofing techniques by excluding *eleven* unknown technologies from the training and development set. The results of the challenge indicates that generalization in spoofing is the most significant problem.

The implementation of SSL models has demonstrated significant performance gain for image, video and audio deepfake detection [23, 24, 20, 25]. However, manipulated content en-

[1] https://github.com/topics/deepfake

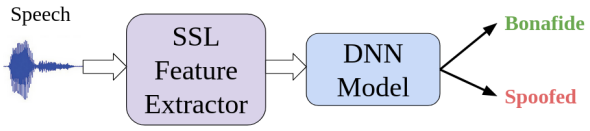

Figure 1: The general framework of proposed deepfake detection systems

countered during testing typically comes from previously unseen datasets or when generated using unknown methods. Notably, performance of detection systems decreases when there is a distinctive difference between the training and test data. Therefore, it is crucial for such systems to be able to generalise to unseen data to increase the robustness against spoofing attacks. To understand the generalisation capability of SSL models, we propose to investigate the SSL models with a multi-stage training framework and conduct cross-dataset evaluations. Our experimentation explored sensitivity of systems towards training datasets and their impact on generalization underlining crucial role of spoofed datasets, ASVSpoof 2019 [26] in comparison with ASVSpoof 2024 [27]. Furthermore, we also propose to use Generative Flow as a way to normalize deepfake embeddings as an augmentation to general framework of deepfake detection. Our experiments show that the ensemble approach significantly enhances generalization across both familiar and out-of-domain scenarios, highlighting the importance of model diversity in overcoming individual system's limitations.

2. Proposed Work

In this section, we outline the proposed framework for the audio deepfake detection system. We begin by briefly describing the three different SSL models used for feature extraction. Next, we discuss the Deep Neural Network (DNN) models and Generative Flow utilized for the classification task. Finally, we detail the end-to-end deepfake detection system for a given speech input.

2.1. SSL models

This section describes the SSL models used as a front-end feature extractor from raw speech waveform. As per the guidelines of the ASVspoof 2024 challenge, SSL models pre-trained on the LibriSpeech dataset were allowed, therefore we opted for WavLM-Base, Wav2Vec2-Large, and HuBERT-Base for feature extraction.

2.1.1. WavLM-Base

The WavLM-base model, hosted on Hugging Face by Microsoft [2], is a pre-trained self-supervised speech model designed to handle a wide range of speech processing applications. It is built on the HuBERT [28] framework and pre-trained on 960 hours of 16kHz speech audio from the LibriSpeech-960 hours dataset, emphasizing both content modelling and speaker identity preservation. The model excels in tasks like speech recognition, classification, and speaker verification, requiring fine-tuning in the supervised learning setting. It is particularly noted for its performance on the SUPERB benchmark [29], showcasing its versatility and effectiveness. The WavLM model integrates masked speech prediction with denoising during pre-training. This dual approach maintains the model's ability to capture speech content while enhancing its performance on non-Automatic Speech Recognition (ASR) tasks through effective denoising. The WavLM-base variant incorporates 12 Transformer encoder layers, each with 768-dimensional hidden states and 8 attention heads, resulting in a total of 94.70 million parameters.

2.1.2. Wav2Vec2-Large

The Wav2Vec2-large-960h model by Facebook, available on Hugging Face [3], is a large pre-trained speech recognition model fine-tuned on 960 hours of Librispeech data sampled at 16kHz. Wav2Vec2 utilizes SSL, which masks speech signals in the latent space and trains them for a contrastive task over quantized latent representations [30]. This enables Wav2Vec2 to perform better on ASR tasks compared to semi-supervised approaches under limited labelled data conditions. Thus illustrating efficiency and effectiveness in speech recognition tasks. Wav2Vec2 is based on transformer networks [31]; we used the Wav2Vec2-large variant, which contains 24 transformer blocks with hidden output dimensions of 1024 and 16 attention heads.

2.1.3. HuBERT-Base

The HuBERT-base (Hidden-Unit BERT) model [28], available on Hugging Face [4], is a SSL model pre-trained on the LibriSpeech-960 dataset, consisting of 960 hours of 16kHz speech. It addresses speech representation challenges by using offline clustering to create aligned target labels for a BERT-like prediction loss. It focuses on masked regions to learn combined acoustic and language models. HuBERT achieves competitive SOTA performance on Librispeech and Libri-light benchmarks over Wav2Vec2, demonstrating significant improvements in word error rate on challenging evaluation subsets. The HuBERT-base variant consists of 12 transformer layers, each with 768 hidden units and 8 attention heads, amounting to approximately 95 million model parameters.

2.2. DNN Models

We have used two DNN models as a backend classifier for audio deepfake detection tasks, namely ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation) [32] and AASIST (Audio Anti-Spoofing Integrated Spectro-Temporal graph attention network) [33], where both DNN approaches have shown SOTA results on antispoofing tasks.

The ECAPA-TDNN architecture, presented by Desplanques et al. in 2020 [32], enhances the x-vector model by extending the temporal context beyond 15 frames using Squeeze-and-Excitation blocks along with attentive statistical pooling. In this work, we opted for the ECAPA-TDNN variant with the convolutional layers filter set to 512 and the output embedding dimension to 192. The AASIST [33] extends RawGAT-ST [34], by incorporating significant improvements such as a heterogeneous stacking graph attention layer [35], a max graph operation for artifact selection, and an updated readout scheme. For this study, we chose to use the lightweight variant of AASIST, as detailed in [33].

[2] https://huggingface.co/microsoft/wavlm-base

[3] https://huggingface.co/facebook/wav2vec2-large-960h

[4] https://huggingface.co/facebook/hubert-base-ls960

2.3. Generative Flow

Generative Flow (Glow) [36], a deep generative model used for tasks such as image synthesis [36], text-to-speech [37], and vocoding [38], maps complex distributions to simple, tractable latent space distributions. Glow is designed using sequential cascaded invertible transformations to estimate complex data distributions by computing the change in log probability density at each step. Unlike traditional generative models like Generative Adversarial Networks [39] or Variational Autoencoders [40], which often require complex training procedures and may experience mode collapse, the Glow model allows for exact log-likelihood computation due to its invertible transformations. This results in a more stable and interpretable latent space.

Each step in Glow consists of an activation normalization, followed by an invertible $1x1$ convolution, and then an affine coupling layer. The Glow is optimized based on the log-likelihood criterion as stated in Equation 1, where the first term refers to the latent variable z sampled from a prior distribution $P_\phi(z)$, typically a multivariate Gaussian distribution $P_\phi(z) = \mathcal{N}(z; 0, I)$. The second term represents the change in log densities transitioning from z_{k-1} to z_k through the transformations $f_{\theta k}$ over K steps of Glow.

$$\mathcal{L}_{Glow} = -logP_\phi(z) - \sum_{k=1,z_0=x}^{k=K,z_K=z} log|det\left(\frac{dz_k}{dz_{k-1}}\right)| \quad (1)$$

where, we utilized $K = 3$, Glow steps for transformations with the input data point x being the deepfake embedding from the last layer of a DNN model. Throughout the Glow transformations, the latent space dimension remains identical to the input data dimension.

2.4. DeepFake Detection Framework

This section describes the general framework of the proposed deepfake detection system. First, we extract SSL features from the models described in section 2.1: WavLM-base, Wav2Vec2-Large, and HuBERT-base, using the last hidden layer output as the SSL features. These extracted features are then given as input to the DNN models to classify the given speech as either bonafide or spoofed, as illustrated in Fig. 1. During the training phase, both the DNN models (AASIST or ECAPA-TDNN) and the SSL models are fine-tuned in a supervised learning setting using cross-entropy loss. We propose incorporating the Glow layer as detailed in section 2.3. We augmented a Glow layer to the pre-trained SSL-DNN systems trained on the deepfake detection task as Glow can learn the distribution of artifacts in deepfake speech. The hidden output of the DNN's last layer is passed to Glow, mapping it to a latent variable, which is then fed into a feed-forward layer to classify as either bonafide or spoofed. For systems augmented with the Glow layer, the loss criterion is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cross-entropy} + \beta * \mathcal{L}_{Glow} \quad (2)$$

where, β is cyclic annealing [41] factor to ensure smooth learning using Glow based systems.

3. Experimentation

First, we describe the datasets used for training and evaluating the proposed systems. Next, we state the experimental setup, followed by the evaluation metrics.

3.1. Dataset

We assess the robustness of the systems using evaluation sets from the ASVspoof 2019, 2021, and 2024 datasets, as well as "In-the-Wild" [22] dataset. The ASVspoof 2019 dataset serves as a benchmark for evaluating the robustness of ASV systems against spoofing attacks and includes subsets for LA and Physical Access (PA) [26, 42]. The ASVspoof 2021 dataset [43] builds on previous editions by introducing more sophisticated spoofing techniques and a focus on deepfake audio. The latest ASVspoof 2024 dataset [27] addresses evolving audio spoofing threats with advanced techniques and a broader evaluation framework. We have provided evaluations on both the development and the progress set of ASVspoof 2024 dataset. In Section 4, we provide an in-depth experimentation of the proposed systems trained on the ASVspoof 2024 challenge dataset, as well as on the systems trained with the ASVspoof 2019 dataset.

3.2. Experimental setup

We develop multiple deepfake detection systems by combining various SSL models (described in Section 2.1) with DNN models (Section 2.2) and the Glow layer (Section 2.3). A more detailed description of the different variants of SSL model-based deepfake detection systems is presented in Section 4. All systems were trained with a learning rate of $1e-5$ and used AdamW[5] [44] optimizer with a weight decay of $1e-3$. The training was conducted over 100 epochs, with early stopping at 10 epochs, based on validation loss for most systems. We use the batch size of 8 along with gradient accumulation over 8 batches. The experiments were initialized with a seed value of 42 and conducted on A100 GPUs and a single L40S GPU, with each system requiring approximately 20 hours of training. Additionally, for Glow-based models, an annealing coefficient was multiplied with the Glow loss, where the annealing coefficient β is varied from 0 to 1 in cyclic manner with linear increment over each batch of training.

3.3. Evaluation

We compute the equal error rate (EER) and the minimum value of the Detection Cost Function (minDCF) to evaluate the performance on systems trained using ASVspoof 2019 and 2024 datasets. Additionally, we reported minDCF, actDCF, cost of log-likelihood ratios (Cllr), and EER on systems reported on ASVspoof 2024 track-1 [27].

4. Results and Analysis

This section describes the performance of various SSL-based systems, evaluated on the ASVspoof challenge datasets of 2019 LA, 2021 LA and DF, In-the-Wild, and ASVspoof 2024 Dev set. We present a comprehensive analysis on generalization capabilities and impact of training data on deepfake detection systems. We have investigated the systems performance with and without Glow augmentation. We also discuss the results obtained by combining various systems through score fusion using averaging in our ensemble models. We trained the systems explicitly on the ASVspoof 2019 [26, 42] (Section 4.1) and ASVspoof 2024 [27] (Section 4.2) training datasets under closed-set conditions, utilizing only the training data without any additional data-augmentation or external speech datasets. We conducted a detailed comparison with SOTA deepfake de-

[5] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html

Table 1: Performance on **training dataset of ASVspoof 2019** with EER and minDCF on eval sets as ASVspoof 2019, ASVspoof 2021 LA and DF, In-the-Wild, and development set of ASVspoof 2024, where ECAPA refers to ECAPA-TDNN as a DNN model, and Glow refers to additional Glow layer augmented with ECAPA-TDNN.

		ASVspoof 2019		**ASVspoof 2021**			**In-the-Wild**	**2024 Dev.**	
ID	**Systems**	**EER(%) (LA)**	**minDCF** (LA)	**EER(%) (LA)**	**minDCF** (LA)	**EER(%)** (DF)	**EER(%)**	**EER(%)**	**minDCF**
S1	**WavLM+Ecapa**	0.80	0.022	6.68	0.372	15.94	34.64	12.30	0.346
S2	**HuBERT+Ecapa**	1.05	0.031	12.55	0.5397	13.79	38.92	21.74	0.614
S3	**Wav2Vec2+Ecapa**	38.86	0.418	26.69	0.451	21.10	24.98	9.70	0.201
S4	**WavLM+Ecapa+Glow**	1.71	0.054	8.54	0.393	26.26	32.07	30.40	0.755
S5	**HuBERT+Ecapa+Glow**	1.41	0.042	8.16	0.395	21.12	42.58	31.71	0.848
S6	**Wav2Vec2+Ecapa+Glow**	24.90	0.250	13.49	0.316	13.86	**18.34**	11.63	**0.192**
S7	**Fusion A**	0.60	0.020	0.60	0.305	13.68	23.19	**8.74**	0.202
S8	**Fusion B**	1.10	0.035	1.10	0.330	12.73	19.89	12.15	0.244
S9	**Fusion C**	**0.59**	**0.019**	**0.59**	**0.292**	**11.75**	18.84	8.79	0.201

tection systems namely AASIST [33], RawGAT-ST [34] and Whisper based system [18] in Table 2.

Table 1 presents the performance of SSL-based systems trained on ASVspoof 2019 dataset, and Table 3 presents the evaluation results for the systems trained on the ASVspoof 2024 dataset. The system Fusion A (S7) refers to ensemble score fusion by averaging scores from SSL systems trained without Glow which are S1, S2, and S3. Similarly, Fusion B (S8) refers to the averaging scores as ensemble system from SSL systems with Glow, S4, S5, S6. We showed performance by averaging scores from all systems from S1 to S6 as Fusion C system (S9).

4.1. Trainset: ASVspoof 2019

As evident from Table 1, the WavLM+Ecapa (S1) system achieved an exceptionally low EER of 0.80% and a minDCF of 0.022 on the 2019 LA dataset, indicating high accuracy. However, its performance degraded significantly on the ASVspoof 2021 and In-the-Wild datasets. Wav2Vec2+Ecapa (S3) achieved high EERs across all datasets, indicating poor performance, but performed better in the In-the-Wild scenario and on the ASVspoof 2024 Dev set with an EER of 9.70%. HuBERT+Ecapa (S2) had a low EER on the 2019 LA dataset, but showed degraded performance in other scenarios, especially In-the-Wild. Glow-based systems WavLM (S4) and HuBERT (S5) obtained high EERs across most datasets, struggling particularly in the In-the-Wild scenario. Wav2Vec2+Ecapa+Glow (S6) showed varied performance, achieving the best results among non-fusion systems in the In-the-Wild scenario and relatively better performance on the 2024 Dev set. Fusion A (S7) and Fusion B (S8) both performed remarkably well across all datasets, with Fusion A achieving one of the lowest EERs on 2024 Dev set and Fusion B showing similar performance. Fusion C (S9) demonstrated the best overall performance, with the lowest EERs on the 2019 and 2021 LA datasets, demonstrating extensive capabilities to perform well across all other datasets, including the 2024 Dev set as well as In-the-Wild scenarios.

Table 2 presents comprehensive evaluation and comparison with the SOTA audio deepfake detection systems trained under same setting as on ASVspoof 2019 dataset and evaluated on ASVspoof 2019, 2021 (LA and DF), In-the-Wild and Dev set of ASVspoof 2024, except MesoNet-Whisper-MFCC system [18]. The RawGAT-ST [34] and AASIST [33] systems exhibited a moderate performance with an EER of 1.22%, and 0.83% on the ASVspoof 2019 dataset respectively, but the effectiveness significantly dropped on other evaluation sets. As Whisper-based system is also trained partly on DF ASVSpoof 2021 set, we observed significantly lower EER on the respective dataset. Despite being pre-trained on large amount of speech data, Whisper-based system was unable to generalise well on

Table 2: Performance measured using EER for comparison of proposed systems with state-of-the-art systems **trained on ASVspoof 2019 dataset** (except MesoNet,Whisper,MFCC) and evaluated of proposed systems, where 2019 refers to eval ASVspoof 2019 eval set, 2021 as ASVspoof 2021 and Dev 2024 for development set of 2024

Systems	**2019**	**2021 LA**	**2021 DF**	**In-the-Wild**	**Dev 2024**
RawGAT-ST[34]	1.22	10.23	37.15	52.54	37.8
AASIST[33]	0.83	11.46	21.06	43.01	37.94
MesoNet,Whisper,MFCC[18]	5.83	15.82	0.36	26.72	3.26
WavLM+Ecapa	0.80	6.68	15.94	34.64	12.3
WavLM+Ecapa+Glow	1.71	8.54	26.26	32.07	30.4
S9 Fusion C	**0.59**	**4.65**	**11.75**	**18.84**	**8.79**

datasets other than In-the-Wild, and Dev 2024.

The WavLM-based systems show competitive results with AASIST and RawGAT-ST on ASVSpoof 2019 eval set. Also, the SSL-based systems consistently display improved performances across other evaluation set. Similar to SSL systems, RawGAT-ST and AASIST operates with speech as input. Their pre-training for masked prediction on large-scale datasets highlights the performance generalization. The augmentation of WavLM+Ecapa with Glow resulted in higher EERs of 1.71% and a t-DCF of 0.054 on ASVspoof 2019, but a slight improvement In-the-Wild scenario with an EER of 32.07%. The ensemble S9 system outperformed all other systems, achieving the lowest EER of 0.59% and t-DCF of 0.0193 on the ASVspoof 2019 dataset, and with an EER of 18.84% demonstrating the best performance in the In-the-Wild scenario. This highlights the superior robustness of the ensemble S9 system across diverse evaluation conditions.

4.2. Trainset: ASVspoof 2024

We conducted a similar study with proposed SSL-based systems, as stated in Section 4.1 on the training dataset of ASVspoof 2024, as shown in Table 3. The results indicate that despite better performances by WavLM-based systems with (S1) and without Glow (S4) on the Dev set of ASVspoof 2024, they have shown moderate performance on other evaluation sets. Notably, these systems demonstrated significantly degraded performance in the In-the-Wild scenario, indicating challenges in handling more varied and unpredictable data.

The Wav2Vec2-based systems (S3), including Glow (S6) demonstrated high EERs across all datasets, with the worst performance observed in the In-the-Wild scenario with an EER of 42.05% and an EER of 24.64% on the 2024 Dev set. The HuBERT+Ecapa system (S2) achieved a notably low EER of 2.99% on the 2019 LA dataset but achieved varied results on other datasets, including an EER of 13.13% on the

Table 3: Performance on **training dataset of ASVspoof 2024** with EER and minDCF on eval sets as ASVspoof 2019, ASVspoof 2021 LA and DF, In-the-Wild, and development set of ASVspoof 2024, where ECAPA refers to ECAPA-TDNN as a DNN model, and Glow refers to additional Glow layer augmented with ECAPA-TDNN.

		ASVspoof 2019		**ASVspoof 2021**			**In-the-Wild**	**2024 Dev.**	
ID	**Systems**	**EER(%) (LA)**	**minDCF** (LA)	**EER(%) (LA)**	**minDCF** (LA)	**EER(%)** (DF)	**EER(%)**	**EER(%)**	**minDCF**
S1	**WavLM+Ecapa**	16.67	**0.314**	**14.89**	0.485	16.15	36.33	5.23	0.137
S2	**HuBERT+Ecapa**	**2.99**	0.425	18.88	0.577	17.99	38.56	13.13	0.290
S3	**Wav2Vec2+Ecapa**	34.86	0.812	37.54	0.880	35.07	42.05	24.64	0.618
S4	**WavLM+Ecapa+Glow**	17.65	0.348	16.12	**0.483**	**13.07**	33.13	**2.47**	**0.070**
S5	**HuBERT+Ecapa+Glow**	0.86	0.576	37.57	0.729	27.59	40.42	19.51	0.355
S6	**Wav2Vec2+Ecapa+Glow**	64.49	0.977	64.28	0.971	60.07	69.3	53.67	0.546
S7	**Fusion A**	17.65	0.334	15.35	0.493	15.95	34.43	6.95	0.177
S8	**Fusion B**	20.58	0.398	18.32	0.519	16.59	33.34	6.23	0.139
S9	**Fusion C**	17.87	0.336	15.36	0.484	15.78	**32.81**	5.66	0.141

Table 4: Evaluation of systems **trained using the ASVspoof 2024 dataset** and performance measures on the ASVspoof 2024 progress set.

Systems	**minDCF**	**EER(%)**	**Cllr**	**actDCF**
AS1 WavLM+AASIST	0.328	13.08	0.625	0.342
AS2 HuBERT+AASIST	0.294	14.48	0.812	0.433
AS3 Wav2Vec2+AASIST	0.288	22.30	0.685	0.396
S1 WavLM+Ecapa	0.197	7.86	0.405	0.274
S2 HuBERT+Ecapa	0.279	10.43	0.586	0.641
S3 Wav2Vec2+Ecapa	0.734	30.37	2.537	0.996
S4 WavLM+Ecapa+Glow	**0.199**	**7.11**	**0.289**	**0.202**
S10 Fusion	**0.164**	**6.41**	0.325	0.212

2024 Dev set and 38.56% in the In-the-Wild scenario. The WavLM+Ecapa+Glow system (S4) demonstrated improved performance over its non-Glow counterpart on several datasets, achieving an EER of 2.47% on the 2024 Dev set and 33.13% in the In-the-Wild scenario. The HuBERT+Ecapa+Glow (S5) achieved the lowest EER of 0.86% on the 2019 LA dataset but underperformed on the 2021 LA dataset and in the In-the-Wild scenario, with EERs of 37.57% and 40.42%, respectively. The Wav2Vec2+Ecapa+Glow system (S6) achieved degraded performance across all datasets, with high EERs of 64.49% particularly on the 2019 LA dataset and 53.67% on the 2024 Dev set.

Among the fusion systems, Fusion A (S7) shows competitive performance with an EER of 17.65% on the 2019 LA dataset and an EER of 6.95% on the 2024 Dev set, but a higher EER of 34.43% in the In-the-Wild scenario. Fusion B (S8) had similar results with an EER of 20.58% on the 2019 LA dataset and 6.23% on the 2024 Dev set, while Fusion C (S9) delivered consistent performance with an EER of 17.87% on the 2019 LA dataset and 5.66% on the 2024 Dev set, and the best performance in the In-the-Wild scenario among fusion systems with an EER of 32.81%. While individual systems showed strengths on specific datasets, fusion systems, particularly Fusion C, demonstrated more balanced performance across diverse conditions, indicating the potential benefits of combining multiple systems for robust audio deepfake detection.

The evaluation results of various systems trained on the ASVspoof 2024 dataset, using the ASVspoof 2024 progress set, are summarized in Table 4. Metrics used include minDCF, EER, Cllr, and actDCF. Among the AASIST-based systems, AS1 (WavLM+AASIST) showed the best performance with a minDCF of 0.328 and an EER of 13.08%, outperforming AS2 (Wav2Vec2+AASIST) and AS3 (HuBERT+AASIST) in EER. The ECAPA-TDNN-based systems demonstrated notable improvements with S1 (WavLM+Ecapa) achieving a minDCF of 0.197 and an EER of 7.86%. The system S3 (HuBERT+Ecapa) also performed well, while S2 (Wav2Vec2+Ecapa) with an EER of 30.37% and a minDCF of 0.734 demonstrated poor performance. The addition of Glow in S4 (WavLM+Ecapa+Glow) further enhanced the performance, achieving a minDCF of 0.199 and the lowest EER among single systems with an EER of 7.11%, with improvements in Cllr and actDCF. It is worth mentioning that among various SSL systems, WavLM shown better suitability to audio deepfake detection as a downstream task, even though all three SSL systems are pre-trained on same amount of speech data. Furthermore, on the progress set, fusion system (S10) is calibrated as weighted sum of scores from 4 best performing systems on progress set as an ensemble system, with formulation as given below and $\mathcal{S}$ indicating scores of a given system,

Table 5: Evaluation of systems **trained using the ASVspoof 2024 dataset** and ASVspoof 2024 as the evaluation set.

Systems	**minDCF**	**actDCF**	**Cllr**	**EER%**
RawNet2[27]	0.827	0.992	4.094	36.04
AASIST[27]	0.711	0.93	4.001	29.12
S10 Fusion	**0.321**	**0.371**	**0.581**	**11.24**

$$\mathcal{S}_{S10} = 0.66 * \mathcal{S}_{S1} + 0.16 * \mathcal{S}_{S2} + 0.16 * \mathcal{S}_{AS3} + \mathcal{S}_{S4} \quad (3)$$

The fusion system (S10), combining multiple system's score, achieved the best overall performance with a minDCF of 0.164 and the lowest EER at 6.41%, highlighting the effectiveness of fusion on progress set. Although S10 maintained competitive Cllr and actDCF values, S4 showed better performance in these metrics. Overall, the fusion systems and the combination of WavLM with Ecapa and Glow demonstrated the most balanced and effective performance in the audio deepfake detection task, highlighting the advantages of integrating multiple detection approaches to enhance overall system efficacy.

Table 5 highlights the substantial benefits of system fusion in audio deepfake detection. While individual systems like RawNet2 and AASIST shown with higher error rates and poorer cost function values. The fusion system (S10) leverages the strengths of multiple systems to achieve superior performance as a mitigation of out-of-domain scenarios. This suggests that diverse model architectures contribute complementary strengths, reducing the likelihood of detection errors and improving robustness against various deepfake techniques. In the context of audio deepfake detection, where the complexity and variability of attacks can be high, a fusion approach appears to be particularly advantageous. By integrating different systems, the fusion system can better generalize across different types of audio manipulations, providing a more reliable defense mechanism. This finding underscores the importance of continued research into ensemble methods and the development of fusion strategies.

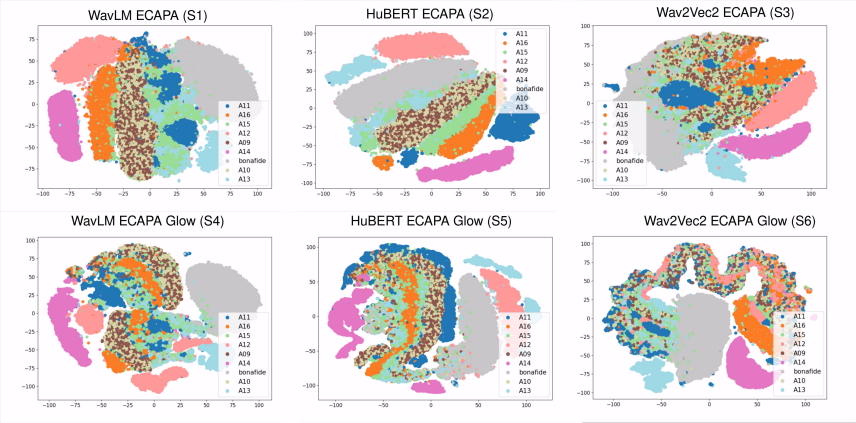

Figure 2: The t-SNE visualization of deepfake embeddings extracted on Dev set of ASVSpoof 2024 and systems, S1 to S6 trained on ASVSpoof 2024 dataset, where A09, ToucanTTS; A10, A09+HifiGANv2; A11, Tacotron2; A12, In-house unit-select; A13, StarGANv2-VC; A14, YourTTS; A15, VAE-GAN; A16, In-house ASR-based [27].

5. Discussion

The evaluation results indicate several key observations regarding the performance of different systems under various evaluation set conditions. We observed similar performance patterns in Tables 1 and 2, where WavLM and HuBERT-based systems consistently outperformed Wav2Vec2-based systems, despite having smaller model parameter sizes. This highlights the efficiency and effectiveness of WavLM and HuBERT models in this task. Notably, WavLM is built upon the HuBERT framework with additional joint speech denoising along with mask prediction, which differentiates it from both HuBERT and Wav2Vec2 systems [45].

The addition of Glow to the WavLM model did not bring about the anticipated improvements in performance on the ASVSpoof 2019 train set. While Glow-augmented systems have shown enhancements in certain contexts, such as in the 2024 dataset for WavLM, this improvement was not observed uniformly. This suggests that the benefits of Glow may be context-dependent on the feature distribution space. Notably, in Table 3 on the ASVSpoof 2024 trainset, the WavLM+Glow approach exhibited state-of-the-art performance as a single system, even surpassing the fusion system (S10) on Cllr and act-DCF metrics. This indicates that generative models, such as Glow, have potential in normalizing the latent space distribution of deepfake embeddings, leading to improved detection performance. In real audio deepfake detection scenarios, generalization is essential to effectively address the bias and variance problem. A detection system that generalizes well can accurately identify deepfakes across various datasets and attack methods, minimizing overfitting (low bias) and underfitting (low variance). Hence, the balance in score calibration is crucial and illustrated by the performance of the fusion system (S10). Thus, ensuring that the system remains robust and reliable when encountering novel and diverse deepfake techniques in real-world applications.

Figure 2 presents t-SNE visualizations of deepfake embeddings, highlighting how different systems and their combinations with Glow impact the embeddings' structure and discriminative ability regarding attack types. Overall, attack IDs A13 (StarGANv2-VC) and A15 (VAE-GAN) show overlap with bonafide across all systems. These t-SNE plots show that the systems are less effective against voice conversion attacks but perform well against TTS-based attacks. The clustered structures in HuBERT and WavLM systems are comparatively better defined than those in Wav2Vec2, which is consistently reflected in the system performances shown in Table 1 and Table 3. From t-SNE visualization, it shows the potential in using Flow metric learning [46, 47] to achieve clustered deepfake embeddings with generative models. Specifically, metric learning can be applied as an auxiliary loss term in the WavLM+Ecapa+Glow (S3) system. We observed that the performance of the same system varied across two different training datasets. The ASVSpoof 2019 dataset enabled robust performance across all evaluation sets, whereas SSL systems showed poor generalization using ASVSpoof 2024 dataset. This underscores the continued importance of the ASVSpoof 2019 dataset with legacy synthesis attacks in developing audio deepfake detection systems. Additionally, these findings highlight the fundamental challenge of bias and variance in supervised learning for out-of-domain scenarios [48, 49]. Thus, emphasizing the importance of score calibration, regularization, feature selection, and cross-data evaluation in developing systems to handle real-life deepfake attacks.

6. Conclusion

The adaptation of SSL models for deepfake detection has shown promising results over the years. However, their ability to generalize to unseen data remains understudied. To address this, we developed deepfake detection systems using WavLM, HuBERT and Wav2Vec2 as SSL models and augmented them with Glow. Our findings indicated that ensemble systems demonstrate superior generalization across various scenarios compared to single system. The results also highlight the impact of training datasets on system performance, emphasizing the importance of the ASVSpoof 2019 dataset. Pre-training with speech denoising using WavLM showed superior performance compared to HuBERT and Wav2Vec2, showing vital role of pre-training. Additionally, in the ASVSpoof 2024 dataset, using Glow for embedding normalization with the WavLM system showed promising results, underscoring the importance of latent space distribution and disentanglement concerning spoofing artifacts. Extensive evaluation of state-of-the-art systems revealed their limitations under various spoofing scenarios. This study paves the way for future research to enhance the generalization ability of SSL models and establishes a benchmark for the research community to evaluate system performance in out-of-domain deepfake scenarios comprehensively.

7. Acknowledgements

For the second, fourth, and fifth authors, this work was granted access to the HPC/AI resources of IDRIS under the allocation 2023-AD011013889R1, provided by GENCI and funded by the Côtes d'Armor Departmental Council. For the third and sixth authors, work was partially supported by the Swiss National Science Foundation project no 219726 on "Pathological Speech Synthesis" and the Innosuisse flagship project no PFFS-21-47 on "Inclusive Information and Communication Technologies.

8. References

[1] Abdulqader M. Almars, "Deepfakes Detection Techniques Using Deep Learning: A Survey.," *Journal of Computer and Communications*, 2021.

[2] Tao Wang, Ruibo Fu, Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Chunyu Qiang, and Shiming Wang, "Prosody and Voice Factorization for Few-Shot Speaker Adaptation in the Challenge M2voc 2021," in *ICASSP*, 2021.

[3] X. Tian J. Yamagishi R. K. Das T. Kinnunen Z.-H. Ling Z. Yi, W.-C. Huang and T. Toda, "Voice conversion challenge 2020: Intralingual semi-parallel and cross-lingual voice conversion," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020.

[4] Ali Swenson and Will Weissert, "New hampshire investigating fake biden robocall meant to discourage voters ahead of primary," *url:https://apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5*, 2024.

[5] Nicholas Diakopoulos and Deborah Johnson, "Anticipating and addressing the ethical implications of deepfakes in the context of elections," *New Media & Society*, 2021.

[6] Loreben Tuquero, "AI detection tools for audio deepfakes fall short. How 4 tools fare and what we can do instead," *url:https://www.poynter.org/fact-checking/2024/deepfake-detector-tool-artificial-intelligence-how-to-spot/*, 2024.

[7] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie el Khoury, "Generalization of Audio Deepfake Detection," in *The Speaker and Language Recognition Workshop*, 2020.

[8] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, L. Ma, and Yang Liu, "DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[9] Zhenjie Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li, "Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks," in *INTERSPEECH*, 2020.

[10] Alejandro Gomez-Alanis, Jose A. Gonzalez-Lopez, S. Pavankumar Dubagunta, Antonio M. Peinado, and Mathew Magimai.-Doss, "On Joint Optimization of Automatic Speaker Verification and Anti-spoofing in the Embedding Space," *IEEE Transactions on Information Forensics and Security*, 2021.

[11] Alejandro Gomez-Alanis, Antonio M. Peinado, Jose A. Gonzalez, and Angel M. Gomez, "A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 2019.

[12] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao, "Audio Deepfake Detection: A Survey," *ArXiv*, vol. abs/2308.14970, 2023.

[13] Zhizheng Wu, Nicholas W. D. Evans, Tomi H. Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, 2015.

[14] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco, and Héctor Delgado, "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[15] Madhu R. Kamble, Hardik B. Sailor, Hemant A. Patil, and Haizhou Li, "Advances in anti-Spoofing: from the perspective of ASVspoof challenges," *APSIPA Transactions on Signal and Information Processing*, 2020.

[16] Choon Beng Tan, Mohd. Hanafi Ahmad Hijazi, Norazlina Binti Khamis, Puteri Nor Ellyza Nohuddin, Zuraini Zainol, Frans Coenen, and Abdullah Bin Gani, "A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction," *Multimedia Tools and Applications*, 2021.

[17] Aakshi Mittal and Mohit Dua, "Automatic speaker verification systems and spoof detection techniques: review and analysis," *International Journal of Speech Technology*, 2022.

[18] Piotr Kawa, Marcin Plata, Michał Czuba, Piotr Szyma'nski, and Piotr Syga, "Improved DeepFake Detection Using Whisper Features," *INTERSPEECH*, 2023.

[19] Xiaohui Liu, Meng Liu, Longbiao Wang, Kong Aik Lee, Hanyi Zhang, and Jianwu Dang, "Leveraging Positional-Related Local-Global Dependency for Synthetic Speech Detection," in *ICASSP*, 2023.

[20] Nicolas M. Muller, Nicholas Evans, Hemlata Tak, Philip Sperl, and Konstantin Bottinger, "Harder or Different? Understanding Generalization of Audio Deepfake Detection," *ArXiv*, vol. abs/2406.03512, 2024.

[21] Wanying Ge, Xin Wang, Junichi Yamagishi, Massimiliano Todisco, and Nicholas W. D. Evans, "Spoofing attack augmentation: can differently-trained attack models improve generalisation?," *ArXiv*, vol. abs/2309.09586, 2023.

[22] Nicolas Michael Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger, "Does Audio Deepfake Detection Generalize?," *INTERSPEECH*, 2022.

[23] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging Properties in Self-Supervised Vision Transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[24] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang, "Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection," *CVPR*, 2022.

[25] Eros Rosello, Alejandro Gomez-Alanis, Angel Manuel Gómez, and Antonio M. Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *INTERSPEECH*, 2023.

[26] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas W. D. Evans, Md. Sahidullah, Ville Vestman, Tomi H. Kinnunen, Kong Aik LEE, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, and Zhenhua Ling, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech Language*, 2019.

[27] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *ASVspoof Workshop 2024 (accepted)*, 2024.

[28] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel rahman Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[29] Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdel rahman Mohamed, and Hung yi Lee, "SUPERB: Speech processing Universal PERformance Benchmark," in *INTERSPEECH*, 2021.

[30] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: a framework for self-supervised learning of speech representations," in *NIPS*, 2020.

[31] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is All you Need," in *NIPS*, 2017.

[32] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," *INTERSPEECH*, 2020.

[33] Jee weon Jung, Hee-Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong-Jin Lee, Ha jin Yu, and Nicholas W. D. Evans, "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks," *ICASSP*, 2021.

[34] Hemlata Tak, Jee weon Jung, Jose Patino, Madhu R. Kamble, Massimiliano Todisco, and Nicholas W. D. Evans, "End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection," *ASVSpoof Workshop*, 2021.

[35] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, Philip S. Yu, and Yanfang Ye, "Heterogeneous Graph Attention Network," *The World Wide Web Conference*, 2019.

[36] Diederik P. Kingma and Prafulla Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," *NIPS*, 2018.

[37] Ajinkya Kulkarni, Vincent Colotte, and Denis Jouvet, "Analysis of expressivity transfer in non-autoregressive end-to-end multispeaker TTS systems," in *INTERSPEECH*, 2022.

[38] Ryan J. Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," *ICASSP*, 2019.

[39] Gilad Cohen and Raja Giryes, "Generative Adversarial Networks," *ICCCNT*, 2022.

[40] Diederik P. Kingma and Max Welling, "Auto-Encoding Variational Bayes," in *ICLR*, 2014.

[41] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin, "Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing," in *NAACL*, 2019.

[42] Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas W. D. Evans, Tomi H. Kinnunen, and Kong Aik LEE, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *INTERSPEECH*, 2019.

[43] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md. Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik LEE, Tomi H. Kinnunen, Nicholas W. D. Evans, and Héctor Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," *ArXiv*, 2021.

[44] Ilya Loshchilov and Frank Hutter, "Fixing Weight Decay Regularization in Adam," *ArXiv*, vol. abs/1711.05101, 2017.

[45] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Micheal Zeng, and Furu Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, 2021.

[46] Ajinkya Kulkarni, Vincent Colotte, and Denis Jouvet, "Transfer Learning of the Expressivity Using FLOW Metric Learning in Multispeaker Text-to-Speech Synthesis," in *INTERSPEECH*, 2020.

[47] Ajinkya Kulkarni, Vincent Colotte, and Denis Jouvet, "Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis," *29th European Signal Processing Conference (EUSIPCO)*, 2021.

[48] Stuart Geman, Elie Bienenstock, and René Doursat, "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, 1992.

[49] Yehuda Dar, Vidya Muthukumar, and Richard Baraniuk, "A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning," *ArXiv*, vol. abs/2109.02355, 2021.