SoK: The Good, The Bad, and The Unbalanced: Measuring Structural Limitations of Deepfake Media Datasets

Seth Layton, Tyler Tucker, Daniel Olszewski, Kevin Warren, Kevin Butler, and Patrick Traynor
University of Florida

Abstract

Deepfake media represents an important and growing threat not only to computing systems but to society at large. Datasets of image, video, and voice deepfakes are being created to assist researchers in building strong defenses against these emerging threats. However, despite the growing number of datasets and the relative diversity of their samples, little guidance exists to help researchers select datasets and then meaningfully contrast their results against prior efforts. To assist in this process, this paper presents the first systematization of deepfake media. Using traditional anomaly detection datasets as a baseline, we characterize the metrics, generation techniques, and class distributions of existing datasets. Through this process, we discover significant problems impacting the comparability of systems using these datasets, including unaccounted-for heavy class imbalance and reliance upon limited metrics. These observations have a potentially profound impact should such systems be transitioned to practice - as an example, we demonstrate that the widely-viewed best detector applied to a typical call center scenario would result in only 1 out of 333 flagged results being a true positive. To improve reproducibility and future comparisons, we provide a template for reporting results in this space and advocate for the release of model score files such that a wider range of statistics can easily be found and/or calculated. Through this, and our recommendations for improving dataset construction, we provide important steps to move this community forward.

1 Introduction

Deepfake media, known colloquially as deepfakes, are videos, images, and speech that are generated from deep learning models to appear as if they represent genuine snapshots of reality. Whether focused on a specific individual or attempting to create realistic but untargeted humans, these increasingly sophisticated attacks have been enabled by the confluence of powerful GPU hardware and machine learning models. With significant potential for misuse in areas ranging from financial fraud [1] and damage to brands [2] to politics [3], the ability to detect such attacks will become increasingly important. Such a need is already a reality, as some prominent individuals claim that public statements may be deepfakes and not authentic [4].

Researchers attempting to enter this space are confronted with a surprising challenge: determining which datasets they should use to most meaningfully compare against other defenses. We argue that because detecting deepfake media is an instance of the anomaly detection problem, baselines from that many-decades-old field should guide the construction and selection of datasets in this new one. Through this lens, we provide the first systematization of the deepfake media space, the generation techniques used to create samples, metrics to evaluate detectors, and how datasets are constructed. Through this systematization, we observe significant deviations from classical anomaly detection, yielding several challenges and highlighting the need for guidelines of use.

In so doing, we make the following contributions:

- **Categorize Existing Deepfake Media Datasets:** A wide array of deepfake media datasets exist; however, selecting an appropriate dataset is non-trivial and important. We systematize the current space of deepfake media according to their generation techniques, evaluation metrics, and class distribution.
- **Identify Limitations in Deepfake Media Datasets:** Identifying deepfake media is an instance of the anomaly detection problem. As such, we use popular anomaly detection datasets to form a baseline for comparison. From this baseline, we demonstrate significant deviations in underlying assumptions including performance metrics and class distributions. We then show that these differences result in the overstating of detector performance.
- **Provide Best Practices:** We discuss guidelines for both current and future datasets to facilitate more meaningful measurements and comparisons. These include presenting a range of base rates to deal with real-world measurement uncertainty, acknowledgment and justification of dataset imbalances, and characterization using appro-

priate metrics. Additionally, we make available a metric reporting template to facilitate comparison.

Of crucial importance, our systematization demonstrates that recognition of the base-rate fallacy [5] and the use of limited metrics are significant problems in this emerging community. Failing to understand and correct such methodological issues will not only result in overstated performance, but will make the transition to deployed systems more difficult. As such, we believe that our recommendations will not only make the current state of the art more realistic, but that our artifacts will make future work more comparable and replicable.

The remainder of this paper is as follows: Section 2 systematizes deepfake media and uses popular datasets from anomaly detection to establish a baseline; Section 3 measures the impact on stated performance that results from these significant differences; Section 4 includes discussion and practical applications to better contextualize the performance of deepfake media detectors; Section 5 offers recommendations to both dataset designers and researchers attempting to enter this space; and Section 6 provides concluding remarks.

2 Systematization of Deepfake Media

Synthetic media includes forms of altered content, whether by simple alterations (i.e., cheapfakes) or sophisticated generation techniques (i.e., deepfakes). Recent advances in deep learning democratized deepfakes, making their creation inexpensive and automated. Moreover, while cheapfakes demanded considerable time and expertise for modest results, deepfakes produce remarkably authentic content surpassing what previous methods could accomplish.

2.1 Setup

To investigate the state of deepfake media research, we identify three distinct domains: video, image, and speech. For each domain, we collect the most popular datasets based on Google Scholar citations. Then, we analyze each corresponding dataset by generation technique, how the authors report baseline model performance, and the class distribution of fake to real signals. Contextualizing datasets in this way reveals community trends in deepfake media research from a security perspective. Additionally, as deepfake media detection may be classified as anomaly detection at its core, we use the network intrusion detection (NIDS) field and datasets as a baseline for reasoning; as this is a mature and well-established subcommunity of anomaly detection. We gather the suggested metrics and class distributions from 15 widely-cited networking IDS datasets [6–20] surveyed by Ring et al. [21].

2.2 Deepfake Media

To better understand the types of deepfake media we provide a brief overview of the major categories.

Video: Deepfake video samples are not always fully generative, instead they often rely on a source video including a real human subject. From that video, models will perform face and landmark (facial features) detection before sending encoded individual frames into an ML algorithm and decoding the result as an image of the target subject's face. Finally, the model places that image or "mask" over the subject's face and performs a smoothing procedure around the mask's boundary.
Image: Deepfake image generation follows a similar procedure as deepfake video generation; however, deepfake images may be fully generative and retain a high quality as they do not have to produce a consistent temporal result.
Speech: Deepfake speech is any speech not explicitly spoken by a human. Common forms include voice modulation, voice replaying, voice conversion, audio deepfakes, and speech generation. Deepfake speech generation ranges from fully generated, where the input is a specific text to voice conversion which converts the sample of speech from one individual to make it sound like a different individual.

None of the deepfake media described in this section is inherently malicious, however, the potential for adversarial, or non-consensual, uses has spurred the design of datasets for the creation of detection mechanisms by the community.

2.3 Dataset Characterization

We gather a range of the most popular deepfake datasets from the community to assist future researchers when contributing to the space. We classify each dataset into eight categories. We show a condensed version of the systematization of deepfake media in Table 1, where we focus on generation techniques, metrics, and class distributions. However, an in-depth breakdown of each dataset is provided on our companion website (Tables 7, 8, 9, and 10).[1]

Generation Technique: We analyze the spectrum of generation techniques, and we use these techniques to understand trends inter and intra-domain.
Metric(s): We gather all the metrics that are suggested for each dataset, or used to report baseline model performance. We use these metrics and test their efficacy.
Deepfake/Real Ratio: We use the deepfake/real ratio to calculate the assumed base-rate of incidence for deepfakes. This gives a metric to help contextualization for each dataset.
Deepfake Sample Counts: Deepfake sample counts are split into their respective partitions if these partitions exist. We use the deepfake sample counts as the assumed incidence counts for deepfake media in the datasets. If a dataset contains different variations, we aggregate the total number of deepfake samples in all variations.
Real Sample Counts: Real sample counts are split into their respective partitions (e.g., train/test/validation) if these

[1] https://sites.google.com/view/thegoodthebadandtheunbalanced

Systematization			Reference	Generation Technique						Metrics								Class Distribution (Fake Sample %)									
																		Train					Test				
Type	Publicly Available	Has Baseline		Other	GAN	Neural Net	Auto Encoder	Replay	Transformer	EER	Accuracy	AUROC	Other	Recall	Precision	F1-Score	None	0-25%	25-50%	50-75%	75-100%	N/A	0-25%	25-50%	50-75%	75-100%	N/A
Video	✓	✓	[22–27]	4	3	–	–	–	–	1	3	3	1	2	2	–	–	0	0	5	1	0	0	0	5	1	0
		×	[28]	1	1	–	1	–	–	–	–	–	1	1	1	–	–	0	0	0	1	0	0	0	0	0	1
	×	✓	[29–33]	4	3	–	3	–	–	1	3	2	1	–	–	–	–	1	1	0	3	0	1	1	0	3	0
		×	[34, 35]	2	1	1	3	–	–	–	1	2	2	1	–	–	–	0	0	1	0	0	0	0	0	1	0
Image	✓	✓	[25, 36, 37]	3	2	1	2	–	1	–	4	–	–	2	2	1	–	1	0	3	0	0	1	0	3	0	0
		×	[38]	1	–	–	–	–	–	–	1	–	–	–	–	–	–	0	0	1	0	0	0	0	1	0	0
	×	✓	[33, 39, 40]	3	1	–	–	–	–	1	1	2	2	1	1	1	–	0	2	0	1	0	0	2	0	1	0
		×	[41]	1	–	–	–	–	–	–	–	1	–	–	–	–	–	0	0	1	0	0	0	0	1	0	0
Speech	✓	✓	[42–52]	6	3	5	1	7	1	11	1	–	3	–	–	1	–	0	0	2	8	1	0	0	1	9	1
		×	[53–58]	4	1	4	1	–	1	1	1	–	–	–	–	–	4	1	0	2	0	3	1	0	2	0	3
	×	✓	[32, 59–63]	3	1	3	–	1	–	5	1	1	1	1	1	2	–	0	0	4	2	0	0	1	2	2	1
		×	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0
Deepfake Media Total				32	16	14	11	8	3	20	16	11	11	8	7	5	4	3	3	19	16	4	3	4	15	17	6
NIDS	✓	✓	[8, 14]	2	–	–	–	–	–	–	2	–	2	1	1	1	–	2	0	0	0	0	2	0	0	0	0
		×	[9–13, 64]	6	–	–	–	–	–	–	2	2	1	2	2	2	2	5	1	0	0	0	5	1	0	0	0
	×	✓	[7, 15]	2	–	–	–	–	–	–	2	–	1	2	2	1	–	0	2	0	0	0	0	2	0	0	0
		×	[16–20]	5	–	–	–	–	–	–	–	–	–	–	–	–	5	0	2	2	1	0	2	1	2	0	0
NIDS Total				15	0	0	0	0	0	0	6	2	4	5	5	4	7	7	5	2	1	0	9	4	2	0	0

Table 1: A systematization of deepfake datasets. For the class distribution sections, a heatmap indicates the density of datasets that meet the column criteria in the greater group of datasets.

partitions exist. We use the real sample counts as the assumed prevalence for real media in the datasets. If a dataset contains different variations, we aggregate the total number of real samples in all variations.

Has Baseline Model: A dataset requires some form of baseline model to showcase its efficacy and give researchers a bar for comparison. We use this category to highlight any dataset that does not provide this foundational component.

Weighted Citations: We use the collected datasets to showcase the disparity in dataset popularity in the space. To view a dimension of dataset popularity, we collect the total number of citations that accompany a dataset publication. If there are multiple publications for one dataset, we use the combined total citations from each of the publications (e.g., ASVspoof 2019 has two publications [43, 65]).[2] To obtain an equal comparison, we weight the total citations from each publication by the number of days since publication and multiply that by 365 to get the number of citations per year (i.e., $weighted\_citations = \frac{total\_number\_of\_citations}{days\_since\_publication} * 365$).

Year: We use the year to understand the staying power of a dataset and to separate datasets with multiple iterations.

In Table 1, we group datasets by domain, public availability, and baseline model availability. Each column is a category for classification and the number in a cell represents the total number of datasets in that group that meet the requirement. The generation techniques and evaluation metrics are grouped into the most popular themes, where the remaining are classified as "other". The training and testing class distributions are broadly split into five categories with "N/A" meaning no specific class split is suggested for the set. For most of the datasets, there is no training/test split suggested, so we assume the same class distribution in the train and test sets. Additionally, if a dataset contains multiple versions, we use the aggregate total of each class.

[2] We understand that a citation does not necessarily mean usage of the cited dataset, yet this gives a good measure of awareness of each dataset.

2.4 Generation Techniques

Deepfake samples can be generated using one or more approaches (e.g., GANs, convolutional autoencoders). To investigate the prevalence of select techniques in inter-domain and intra-domain datasets, we include every generation technique that contributed to each dataset.

Categorizing Generation Techniques: We observe over a dozen approaches for each domain, ranging from sourcing samples from other datasets to implementing various types of GANs. While early datasets relied on domain-specific generation techniques (e.g., text-to-speech and voice conversion for deepfake speech, and face-swapping apps for deepfake images), newer datasets settle on ML-based generation techniques to produce high-quality samples. We therefore observe a convergence of generation techniques into GANs and autoencoders often used in conjunction with domain-specific tasks (e.g., voice conversion for deepfake speech). Table 1 shows that GANs and NNs substantially dominate the generation space, yet the other category is the largest, suggesting an increasingly complex landscape for generation algorithms. These approaches allow deepfake media generation to compete with increasingly complex models.

2.5 Detection Performance Metrics

Every deepfake media dataset is designed as a platform for training and evaluating classifiers. Accordingly, they generally include a baseline model and compare its performance to existing models in the space. To achieve a direct comparison, however, work in this space needs to report similar metrics. To explore this, we offer a list of every metric included in the deepfake media and network IDS datasets gathered.

Categorizing Metrics: Datasets in each deepfake domain report several metrics: eight for speech and twelve for both image and video. Speech classification reporting is dominated by Equal Error Rate (EER), inspired by the metric's role in the ASVspoof competitions beginning in 2015. While some datasets include additional metrics (e.g., ADD 2023 includes F1-Score), the community standard remains EER.

Unlike the deepfake speech domain, the deepfake image and video domains do not exhibit an overwhelmingly dominant metric. In both domains, the most popular metric is accuracy followed by Area Under the ROC curve (AUROC). However, at least half of the datasets in all domains only report a *single* performance metric. This approach enables comparison between classifiers using that metric, but is generally insufficient in fully contextualizing classifier performance [66].

Furthermore, examining Table 1 and contextualizing all deepfake media datasets shows that EER is the dominant metric. Yet, *no* network IDS datasets analyzed use EER as a metric. Additionally, precision and recall are among the top metrics in NIDS, but see little usage in deepfake datasets.

2.6 Class Distribution

A base-rate for a given class in a population describes the percentage of samples that belong to that class. This simple yet impactful metric contextualizes the performance of a classifier on a dataset. For example, assume that we are interested in the base-rate of deepfakes for all samples in the wild. In this setting, if we observe that 100 samples online are deepfakes compared to a total of 10 million available samples, we can state that the base-rate for deepfakes is 0.001%. If the base-rate is not explicitly stated for a dataset, we must assume that the base-rate was chosen to be the distribution of the classes present in a dataset. Stated concisely, *classifier performance has no context without an accompanying base-rate*.

Categorizing Class Distributions: Unlike traditional anomaly detection, no studies exist to provide a base-rate of deepfake media on the internet. Therefore, even dataset designers aware of the base-rate fallacy [5] (i.e., misinterpreting performance results based on incorrect statistical assumptions) do not have references for constructing their split of fake and real samples. We expand on the base-rate fallacy on our companion website. Nevertheless, we highlight in Table 1 that splits skew toward the fake class in all domains of deepfake media, suggesting that media on the internet is *primarily fake*. Furthermore, the heatmap in Table 1 shows that deepfake media and NIDS have *inverted* class distributions.

To further expand, Figure 1 breaks down the class distribution of each dataset (train/test and dataset versions are split). Of the three deepfake domains, deepfake speech suggests a base-rate of deepfakes at 88%. Additionally, observing the 90th percentile in network IDS and the 10th percentile in the deepfake video and speech datasets shows minimal overlap at the extremes. As such, anomaly detection problems are exercises in *finding a needle in a haystack*; whereas deepfake media datasets such as ASVspoof2021 seemingly ask practitioners to *find hay in a needlestack.*

3 Examining Deepfake Dataset Construction

Figure 1 shows that speech deepfakes have the most skewed class distribution when compared to anomaly detection. To produce the most faithful representation of model performance, we require the following baseline reproducibility criteria [67] to be met for our experiments: the dataset must be publicly available, contain an explicitly defined train and test split, propose an explicitly defined baseline model, and suggest standard comparison metrics. No video or image deepfake dataset meets these requirements. We therefore focus our experiments on the most skewed and reproducible deepfake media category: speech deepfakes. For the remainder of this deep analysis, we focus on speech deepfake datasets to understand the limitations of operating on the *extreme assumption* of a base-rate environment where the total amount of fake samples substantially outweighs the total amount of

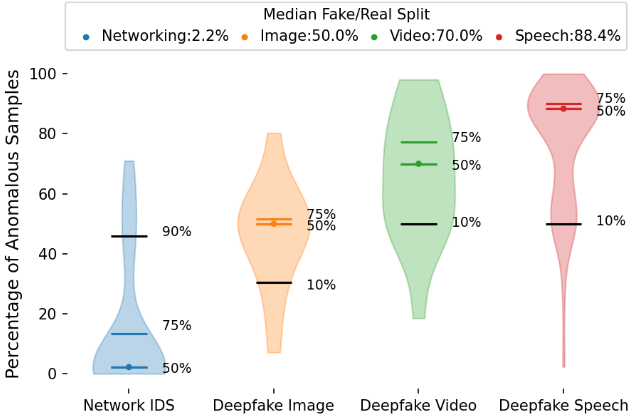


Figure 1: The Ratio of malicious samples to benign samples in 15 network IDS datasets, 23 deepfake speech datasets, 14 deepfake video datasets, and 9 deepfake image datasets. Each group is represented by a violin in the plot above and the quartiles are denoted as the horizontal lines.

real samples. However, in Section 4.1 we show that the issues presented and described here are generalizable.

Deepfake Detection vs. Speaker Verification: In the case of deepfake speech, there are two different categories of datasets: deepfake detection and speaker verification. These are inherently two different tasks, but both employ a form of deepfake speech detection. Speaker verification requires prior knowledge of a specific individual to test if future unseen speech is that specific person. Deepfake detection focuses on whether a sample of speech is *reasonably* human-generated and does not require specific knowledge of any individual. For this paper, we focus on deepfake datasets and detection.

Dataset Down-Selection: There are myriad datasets to use in the space of deepfake speech, therefore the task of selecting the correct dataset is non-trivial. The choice of dataset has a major impact on the final model and results. As such, a filtering process must occur to eliminate datasets that are not useful based on context. One of the most important factors in selecting a dataset is comparability (i.e., if a new technique produces results, is it possible to 1:1 compare that technique with previous results). To showcase this, we filter our list of 23 speech deepfake datasets using 5 simple filters to find suitable datasets. First, we focus on deepfake datasets exclusively to align with the scope of our work. Second, we select publicly available datasets, as comparability only works if a dataset can be used by all. Third, we select datasets that contain an explicit train/test(/validation) split to ensure different models train on the same data. Fourth, we select datasets that have suggested metrics, or use metrics in a custom baseline model, as again for a 1:1 comparison the same metrics must be reported. Finally, we select only datasets that have a comparative baseline model associated. A 1:1 comparison ensures that the community may meaningfully compare results. Without this, it is difficult to compare, or even reproduce results. Table 2 showcases the filter in action; the highlighted rows are the only remaining datasets after applying these five filters.

Dataset Name	**Weighted Citations**	**Train Split**	**Has Metrics**	**Has Baseline**
ASVspoof'21	90.8	✓	✓	✓
CFAD	6.0	✓	✓	✓
Fake or Real	11.2	✓	✓	✗
WaveFake	18.4	✗	✓	✓
In The Wild	14.5	✗	✓	✓
VCC'20	58.5	✗	✗	✗
VCC'18	33.9	✗	✗	✗
VCC'16	28.8	✗	✗	✗

Table 2: All publicly available deepfake speech datasets. The highlighted cells indicate datasets that meet all the following criteria: is a deepfake dataset, is publicly available, has a train/test split, and suggests comparison metric(s).

For the remainder of this paper, we examine the efficacy and value of the ASVspoof 2021 (English) dataset, however, we provide insight into the CFAD (Chinese) dataset where results are interesting or deviate from ASVspoof 2021.

3.1 Setup

We show that ASVspoof and CFAD are the only datasets that meet our set of down-selection criteria. Thus, we examine the ASVspoof 2021 deepfake dataset (i.e., ASV_{trn}, ASV_{dev}, and ASV_{eval}) and the CFAD dataset (i.e., $CFAD_{trn}$, $CFAD_{dev}$, and $CFAD_{eval}$). One minor change we implement is the combination of both test sets in CFAD (test seen and test unseen). As CFAD contains a segment for detecting the generation algorithm used in deepfake speech, the test seen set is defined as samples that were generated with a technique that was used for $CFAD_{trn}$ samples. Conversely, test unseen is defined as samples that were not generated with a technique that was used for $CFAD_{trn}$ samples. The combination of these two test sets comprises the entirety of $CFAD_{eval}$.

In conventional anomaly detection, the positive class represents the most important, the events that should trigger an alarm, and the minority. Despite this, ASVspoof considers real speech as the positive class. To align with anomaly detection, we invert the labels such that deepfake speech is the positive class. We use all baseline models from ASVspoof 2021 (RawNet2, LFCC-LCNN, LFCC-GMM, and CQCC-GMM) and the best two baseline models from CFAD (RawNet2 and LFCC-LCNN) which we define as M_{ASV-RN}, M_{ASV-LL}, M_{ASV-LG}, M_{ASV-CG}, $M_{CFAD-RN}$, and $M_{CFAD-LL}$, respectively. Additionally, we implement the SSL-wav2vec2.0 (M_{ASV-SW}) deepfake detector [68] as it has the best-reported performance, to our knowledge, on ASV_{eval}.

3.2 Research Questions

We create four research questions and conduct experiments to answer each of them.

3.2.1 Reproducibility

The first step in comparison with any model is to verify the reported results. Towards this, we investigate the reproducibility of baseline detection models in our first research question:

RQ1 · Are performance results from baseline detection models on deepfake datasets reasonably reproducible?

We begin by retraining the models mentioned in Section 3.1 to verify their reproducibility which is a growing concern in machine learning [67]. Toward this, we implement each model directly from GitHub without modification [69–71]. We retrain the ASVspoof models using ASV_{trn} and the CFAD models using $CFAD_{trn}$ with default parameters. Using the trained models, we determine per-class probabilities for predictions against ASV_{eval} and $CFAD_{eval}$, respectively.[3] We evaluate the reproducibility of each model by measuring suggested metrics against the reported values of the baseline models.

We show the results of this reproducibility test in Table 3. We show that for 5 of 7 models, our retraining meets or exceeds the reported metrics for that model. The reported EERs for the four ASVspoof baseline models and our measured EERs are all within 4.32% (relative). However, we measured EER for M_{ASV-SW} as 4.14% which is relatively 52.3% worse than the reported 2.85%.

As the CFAD results are reported individually for both test sets (seen/unseen) we must weigh the two results and combine them into a single EER value for comparison to our measured results. We use the formula: $EER = \frac{x*EER_{seen}+(1-x)*EER_{unseen}}{2}$, where x is the ratio of samples in the test seen set and $1-x$ is the ratio of samples in the test unseen set. We show a matched measurement to the reported EER for $M_{CFAD-LL}$. However, we show a substantially better measured EER for $M_{CFAD-RN}$ than what is reported. This highlights an interesting reproducibility issue with our measured results substantially outperforming the reported values. In this case, the disparity between measured and reported favors the authors of the original work, as we do not produce worse results. A deeper investigation into LFCC-LCNN and RawNet2, as these models are used by both datasets, allows an expected value comparison. The similarity of measured and reported EERs for M_{ASV-LL} and M_{ASV-RN} shows that our measured value of $M_{CFAD-RN}$ is similar to $M_{CFAD-LL}$, which leaves the reported value of 23.9% EER for $M_{CFAD-RN}$ as the outlier. This outlier, and the measured M_{ASV-SW} results, suggest that reproducibility in the space may be an issue **(RQ1)**.

Model	Type	EER	TPR	FPR	Sample Prediction Cosine Similarity
M_{ASV-CG}	M	25.4%	44.9%	13.1%	0.195
	R	25.3%	−−	−−	−−
M_{ASV-LG}	M	25.5%	44.2%	8.80%	0.206
	R	25.6%	−−	−−	−−
M_{ASV-LL}	M	22.9%	94.7%	41.7%	0.468
	R	23.5%	−−	−−	−−
M_{ASV-RN}	M	22.1%	95.9%	36.2%	0.551
	R	22.4%	−−	−−	−−
M_{ASV-SW}	M	4.14%	99.9%	26.9%	−−
	R	2.85%	−−	−−	−−
$M_{CFAD-LL}$	M	9.36%	90.9%	10.8%	−−
	R	9.69%	−−	−−	−−
$M_{CFAD-RN}$	M	10.9%	88.4%	9.49%	0.379
	R	23.9%	−−	−−	−−

Table 3: Measured vs. Reported (M/R) reproducibility results from retraining the 5 ASVspoof2021 and 2 CFAD models. All models are retrained and tested against the original train and test sets. Reported values labeled as −− are not provided or derivable from reported metrics.

For the remainder of the paper, we default to using our measured results instead of reported metrics.

3.2.2 Efficacy of Reported Metrics

We hypothesize that current deepfake datasets are *not* sufficiently robust to allow for meaningful evaluation of detection mechanisms. To investigate this claim, we define our second research question:

RQ2 · Do limited, and often singular, metrics as the sole performance measure sufficiently represent the behavior of deepfake detection models?

The main metrics used in the deepfake detection space are Equal Error Rate (EER), Accuracy, Precision, Recall, F1-Score, and AUROC. We show in Table 9 on our companion website that EER is the substantial majority with ~90% of speech deepfake datasets using EER and ~50% exclusively implementing EER, including ASVspoof 2021 and CFAD.

The efficacy of single performance metrics (e.g., EER) has been shown to be inherently flawed. For example, prior work [66] shows that since EER characterizes a family of ROC curves, the performance of a model may change significantly between the possible prediction thresholds, effectively obfuscating results. While ASVspoof 2021 acknowledges that EER is considered deprecated by the ISO/IEC standards [72], EER remains the only universal metric for speech deepfakes.

To understand the meaningfulness of EER, and verify the inability of this limited metric to report performance, we calculate the true positive and false positive rates, and EER for all

[3] For each model (except $M_{ASV-RN}/M_{CFAD-RN}$) we apply a sigmoid function to the model's scoring function output. $M_{ASV-RN}/M_{CFAD-RN}$ already calculate per-class probabilities and no additional processing is necessary.

models defined in Section 3.1. We compare the foundational metrics of each model to investigate whether performance is obfuscated by using EER as a comparison metric.

Table 3 shows that models with similar EER values may have vastly different metrics (i.e., TPR and FPR). Inspecting M_{ASV-LG} and M_{ASV-LL} (highlighted rows in Table 3) which have similar EER values of 25.5% and 22.9% highlights this issue. M_{ASV-LG} has a TPR of 44.2% whereas M_{ASV-LL} has a substantially better TPR of 94.7%; however, M_{ASV-LG} with an FPR of 8.80% considerably outperforms M_{ASV-LL} with an FPR of 41.7%. Thus, similar EER values do not necessarily represent the underlying performance of a model.

Additionally, we examine the predictions from the five baseline ASVspoof 2021 models on each ASV_{eval} sample to understand how models with similar EERs classify samples. As a broad example, two models exhibiting 90% accuracy over a test set, do not necessarily incorrectly predict the same 10% of samples. For the ASVspoof 2021 baseline models, we measure the total number of samples that the models agreed upon. All five models agree on 34.1% of samples, at least 4/5 models agree on 54.9% of samples, and at least 3/5 models agree on 100% of samples. We show that it is relatively rare for all five models to agree on a specific sample. To further highlight the issue, we take the two baseline models with the closest EER values (i.e., M_{ASV-LG} and M_{ASV-CG}) and calculate the cosine similarity between the list of predictions (ordered based on predicted sample name). The cosine similarity for these two models is 0.831, indicating that two models with an EER difference of 0.1% predict differently on individual samples. Furthermore, we select the "best" model from each dataset as a baseline and compute the cosine similarity between it and the other models for that dataset. We show in the right-most column in Table 3 that there is little agreement between these models.

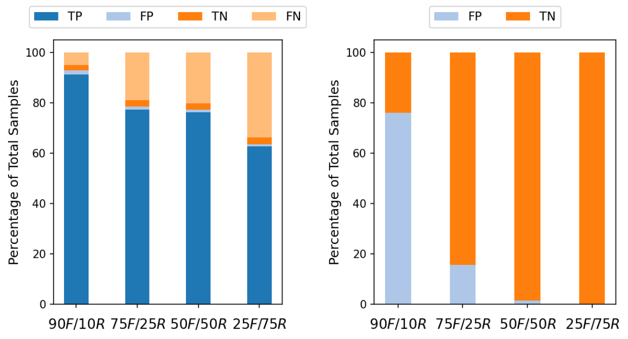
We show that the most used comparison metric (EER) does not honestly report the results of a model and thus, masks intrinsically different performance profiles **(RQ2)**.

3.2.3 Class Distribution in Datasets

Class imbalance in training and evaluation sets may obscure experimental results [73], which exacerbates the misrepresentation caused by unidimensional performance metrics. This leads us to define our third research question:

RQ3 · Does the composition of current deepfake datasets bias classification results?

To examine whether the distribution of classes can create unintentionally biased detectors, we propose two methods to test bias based on $ASV_{trn}/CFAD_{trn}$ and $ASV_{eval}/CFAD_{eval}$. Prior research [74] has shown that such imbalances exist in other datasets and create bias, thus it is necessary to measure whether bias impacts these datasets. Both tests vary the ratio of deepfake to real speech in the training set to determine the effect of class distribution changes on two evaluation sets.



(a) M_{ASV-LL}, test set ASV_{eval}. (b) M_{ASV-LL}, test set RO_{asv}.

Figure 2: To show bias in the training and test sets of deepfake datasets, we examine M_{ASV-LL} with four different training distributions. Note the reduction in true positives, and consequently, the reduction in false positives. Results for M_{ASV-RN}, M_{ASV-SW}, $M_{CFAD-RN}$, and $M_{CFAD-LL}$ may be found on our companion website, Figure 10 and Figure 11.

- **[Test 1]** Tests bias in the evaluation set and varies the training data by augmenting ASV_{trn} with additional real samples and tests against ASV_{eval}. Additionally, we repeat this with $CFAD_{trn}$ against $CFAD_{eval}$.
- **[Test 2]** Tests bias in the training set using the varied models from **[Test 1]** and replaces the evaluation set with entirely real-speech sets RO_{asv} or RO_{cfad}.

We adopt the notation $D_{xF/yR}$ for class distribution, where x/y is the ratio of fake to real speech.

Test 1: ASV_{eval} contains approximately 97% deepfakes and is thus highly imbalanced towards the fake class. To understand any biases inherent in this composition, we retrain the two best baseline models (M_{ASV-RN} and M_{ASV-LL}) and the overall best model (M_{ASV-SW}) with varying class distributions.[4] We test the following class distributions: $D_{90F/10R}$, $D_{75F/25R}$, $D_{50F/50R}$, and $D_{25F/75R}$. We then evaluate each model against the default ASV_{eval} set. To avoid bias by downsampling the overrepresented deepfake class and thus reducing the total number of training samples, we augment the number of real samples in ASV_{trn}. We sub-sample speech from LibriSpeech train-clean-100 [75] to match the distribution of ASV_{trn}.

$CFAD_{eval}$ contains 66% deepfake samples and is not as imbalanced as ASV_{eval}, but is still skewed towards more deepfakes. We retrain the two best baseline models from CFAD ($M_{CFAD-RN}$ and $M_{CFAD-LL}$) with the same training class distributions as ASVspoof.[5]. We then evaluate each model against the default $CFAD_{eval}$ set. We augment the real

[4] The default parameters and epochs are unaltered for each model and training class distribution combination for all datasets.

[5] We downsample the deepfake speech class for $D_{90F/10R}$ and $D_{75F/25R}$

samples in $CFAD_{trn}$ with real-only Chinese speech from the WeNetSpeech Podcast Train set [76] to match distributions.

Specifically, for the real-speech subset of each dataset, we match the sample durations, number of speakers, number of utterances per speaker, and the ratio of female to male speakers. Section 2 on our companion website gives an in-depth breakdown of the augmentation technique and shows the length distribution of the real samples in ASV_{trn}/Librispeech and $CFAD_{trn}$/WeNetSpeech.

Figure 2a shows the true positives, false positives, true negatives, and false negatives for M_{ASV-LL} trained with four different training distributions. For space concerns we only show the results of one model, however, the trends follow for all models and are shown on our companion website in Figures 10 and 11. Additionally, Table 4 on our companion website shows metrics for each combination of training distributions and evaluation sets.

We show that as the training set class distribution moves further away from the testing distribution (i.e., a right move on the x-axis in Figure 2) the overall model performance decreases. To be precise, increasing the number of real speech samples in the training set does not independently improve the ability of the model to predict real speech samples, it also reduces the performance on deepfake speech.

With a simple shift in training distributions, a model's performance changes on the same test set. To put it concisely, a model must train with a similar class distribution to the evaluation set to have performant results. We show that the test sets ASV_{eval} and $CFAD_{eval}$ both contain bias and overstate the performance of a model.

Test 2: To test bias in the training sets ASV_{trn} and $CFAD_{trn}$, we explore a scenario where there are *no* deepfakes in an evaluation set. As the ASV_{eval} set only contains approximately 3% real-class samples, and oversampling or augmenting these samples may add bias to this test, we collect almost 150,000 samples from four publicly available real-speech corpora (TIMIT [77], LJSpeech [78], World English Bible [79], and LibriSpeech train-clean-360 [75]) and label the resulting dataset RO_{asv}. Similarly, we collect real-only Chinese speech samples from WeNetSpeech Meeting Test [76] and label the resulting dataset RO_{cfad}. We use the samples in these real-only evaluation sets to simulate different class distributions than ASV_{eval} and $CFAD_{eval}$ to investigate bias in training.

Figure 2b shows the false positives and true negatives for M_{ASV-LL} with the 4 different training distributions. For brevity, we only show the results of one model, however, the trends follow for all models. See our companion website for the results of other models. This demonstrates the overwhelming number of false positives for $D_{90F/10R}$ in an all-real evaluation set; in fact, all models have a higher false positive rate than a true negative rate. Moving from $D_{90F/10R}$ to $D_{25F/75R}$ dramatically reduces false positive predictions. This further exemplifies the bias of the ASV_{trn} and $CFAD_{trn}$ sets. Models trained on $D_{90F/10R}$ are only effective on a test set with a similar class distribution. Adding a minor amount of the underrepresented class when training reduces bias.

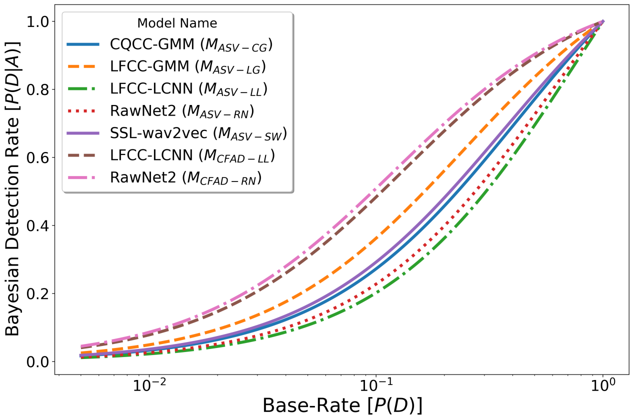


Figure 3: We calculate BDR as a function of the base-rate for all baseline models using the respective test set. Base-rates range from 0.01% (i.e., an approximation based on network IDS measurements), to 100% (i.e., every sample is a deepfake). Referencing the five ASVspoof models in the context of BDR, we observe that M_{ASV-LG} performs the best at every base-rate threshold despite having a substantially worse EER than M_{ASV-LG} (25.5% vs. 4.14%).

We evaluate bias in the ASVspoof 2021 and CFAD train and test datasets with two tests (i.e., changing the underlying training class distributions, and testing the baseline models against an all-real speech set). We show that due to the class distributions, train and test sets impose a bias on the performance of detectors built on these datasets **(RQ3)**.

3.2.4 Characterizing Model Efficacy

As stated previously, classifier performance has no context without an accompanying base-rate. Despite this, no deepfake dataset provides a base-rate to contextualize model results. With this, we define our final research question:

RQ4 · How do current deepfake datasets perform in a base-rate-aware environment?

Abstracting away from the dataset and the class distribution in each set, we examine how framing the output of any given model within the context of a base-rate tests the robustness of that model. Contextualizing a model in the scope of a specific base-rate does not change the underlying performance of that model on any specific evaluation set; however, it may expose the limitations of that performance. We calculate the true and false positive rates for all models retrained in Section 3.2.1. Then, we calculate the Bayesian Detection Rate (BDR) as a function of the base-rate. BDR is defined as:

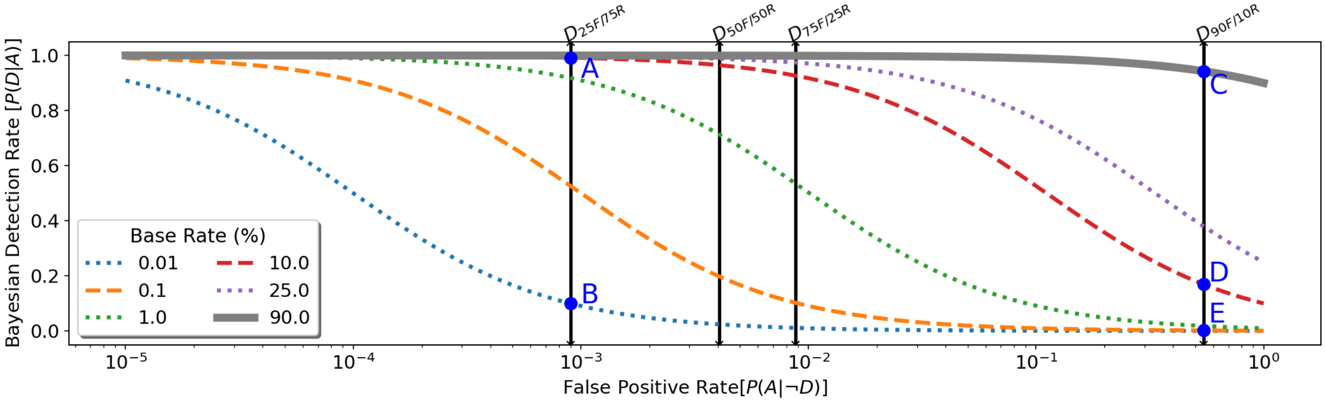


Figure 4: We calculate RawNet2's BDR as a function of the base-rate where the vertical lines represent the measured false positive rates for each class distribution against RO_{asv}. The intersections show where a class distribution maps to a BDR at different base-rates. This mapping gives a visual representation of how the base-rate impacts performance in a real-world scenario, effectively allowing the selection of the best model based on a use case.

$$BDR = P(D|A) = \frac{P(D) \cdot P(A|D)}{P(D) \cdot P(A|D) + P(\neg D) \cdot P(A|\neg D)}, \quad (1)$$

where D is a sample of deepfake media, A is an alarm, $P(D)$ is the probability of encountering a sample of deepfake media, $P(A)$ is the probability of an alarm, $P(D|A)$ is the probability of a sample of deepfake media given an alarm, and $P(A|D)$ is the probability of an alarm given a sample of deepfake media. We provide an in-depth breakdown of BDR on our companion website.

As neither ASVspoof 2021 nor CFAD explicitly state any base-rate assumptions and the base-rate for deepfake media in the wild is unknown, we use a range of base-rates (90%, 75%, 50%, 25%, 10%, 1%, 0.1%, 0.01%). This simulates how well a specific model performs in multiple scenarios.

Figure 3 shows the BDR for the seven baseline models trained with the original training sets at varying base-rates. Referring back to Table 3 shows that M_{ASV-SW} has an EER value that is 83.8% better than M_{ASV-LG}; however, Figure 3 highlights that the BDR of M_{ASV-SW} is strictly worse than M_{ASV-LG} *at every threshold*. Additionally, as the base-rate approaches either extreme, the BDR of every model converges.

Furthermore, to showcase how the BDR helps evaluate the performance of a model, we examine M_{ASV-RN} against the real-only test set without loss of generality, as the training class distribution changes (see Table 4 on our companion website). For M_{ASV-RN}, the correct classification of true negatives increases from 45.60% to 99.91% and the false positive classifications decrease from 54.40% to 0.09%. This is a relative false positive decrease of 99.83% and suggests substantial performance improvements. However, these results do not contextualize the output of the detector within the scope of any base-rate and, as such, overstates the performance of the model. To understand this overstatement, we plot the training class distribution for M_{ASV-RN} against a range of base-rates in Figure 4.[6] For example, point A shows M_{ASV-RN} trained with $D_{25F/75R}$ operating in an environment where 10% of samples are deepfakes. For all base-rates above 10% at point A (i.e., base-rate lines 25% and 90%), nearly all predictions of fake speech made by the model are truly fake samples. Shifting to point B, we show the same model and training distribution as point A, in an environment where the expected encounter with deepfakes is 0.01%. Predictions of deepfake speech made by the model at point B have a 10% probability of truly being fake. Following down to point E shows the original $D_{90F/10R}$ M_{ASV-RN} has a nearly 0% probability of being correct when predicting fake samples. Additionally, point C is the ASVspoof base-rate.

Finally, we evaluate the impact of threshold manipulation on the performance of a model with respect to BDR – as taking the standard 0.5 probabilistic threshold for predictions may not result in the most robust model given a specific context. To achieve this, we take M_{ASV-RN} trained and tested on the original ASV_{trn}/ASV_{eval} sets and set the predictive threshold such that the false positive rate is 1%. Using the default predictive threshold of 0.5, M_{ASV-RN} has a false positive rate of 36.2% and a true positive rate of 95.9%. If we change the threshold to 0.99, the false positive rate drops to 0.99% and the true positive rate drops to 53%. Calculating the BDR from the 0.5 and 0.99 thresholds gives 0.26% and 4.9%, which shows that the model with the better false positive rate has a better BDR (i.e., 20x better). This highlights

[6] The true positive rate is undefined as there are no deepfakes in the RO_{asv} set and we assume a perfectly accurate 1.0 detection rate.

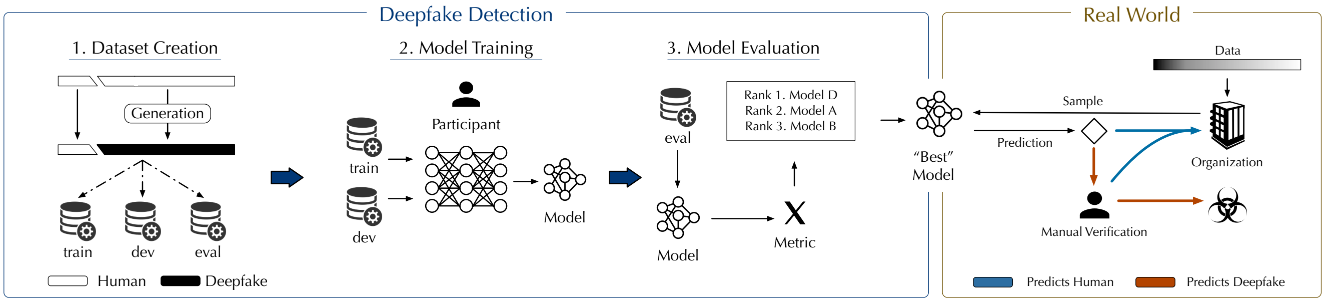


Figure 5: A generalized ecosystem for deepfake dataset creation, including detector benchmarking and a scenario where a detector is used in the real world. To start, a dataset is created and split into training, development, and evaluation subsets (1). These subsets are provided and, using the training set, algorithms are produced based on their lowest metric on the development set (2). The final algorithms are ranked based on their metric over the evaluation set (3). A model may then be selected by an organization based on the ranking. In the real world, all samples classified as a deepfake must be manually reviewed so as not to allow false positives to impact operations and user experience.

the importance of the false positive rate on the BDR and the trade-off (i.e., reduction in the true positive rate) that may be required to achieve this improvement. We show that setting thresholds such as keeping the false positive rate at 1% improves BDR, but the trade-offs must be acknowledged.

We show that every model tested is susceptible to degrading BDR performance when contextualized with a decreasing base-rate and that BDR is mostly affected by the false positive rate. This contextualization may have substantial performance implications as we observe in Figure 4 **(RQ4)**.

3.2.5 Summary of Results

Through our experiments on ASVspoof 2021 and CFAD, we demonstrate the limitations of current deepfake datasets. We show that reproducibility may be an issue with comparisons **(RQ1)**, single metrics (EER), and class imbalance lead to overstated model performance **(RQ2, RQ3)**, and base-rate contextualization (currently undefined for deepfakes) is necessary to meaningfully evaluate a model **(RQ4)**.

4 Discussion

4.1 Larger than Speech Deepfakes

While we focus on speech deepfake datasets, the issues presented in this paper generalize to a greater issue. We do not examine the inherent properties of the data present in any dataset, instead, *we demonstrate that the overall structures inherently lead to misinterpreting results*. While the speech deepfake datasets are substantially more skewed towards the fake class than the video and image datasets, the same issues reported will impact these datasets as well. To demonstrate this, we explore the research questions defined in Section 3.2 on a sample dataset from the image deepfake domain.

There are no image or video deepfake datasets that meet the reproducibility criteria defined in Section 3. However, CIFAKE [37] is only missing concrete baseline model implementation details. As such, we make general assumptions about the missing hyperparameter details of the baseline model. We note that this is not a direct implementation of the baseline system of CIFAKE; but rather a "best-effort" re-implementation. We show (in a detailed analysis on our companion website) that the trends outlined in Section 3.2 follow for the image deepfake domain. While reproducibility is strongly hindered by a lack of explicit documentation (**RQ1**), training and test distributions impose a bias on the performance of the model (**RQ3**), and contextualizing with an appropriate base-rate degrades model performance (**RQ4**).

4.2 Transition to Practice

Deepfake media datasets are collections of human-originated and synthetically generated videos/images/speech for use in machine learning with the explicit task of discrimination between real/fake samples. Unlike other datasets within Security (e.g., network traffic), multiple deepfakes can be generated from one seed input. Thus, class imbalance is an inherent problem within deepfake dataset creation. These datasets enable disparate algorithms and models to benchmark their performance based on specific metrics. This comparison allows for an academically created algorithm to be adopted into practice. Figure 5 shows the general structure of dataset creation, benchmarking, and transitioning to practice. Motivated by this pipeline, we show two practical real-world scenarios.

4.2.1 Speech Deepfakes: A Call Center

To visualize the current issues of speech deepfake datasets, we examine parallel real-world scenarios (SCN_A and SCN_B). Each scenario is a call center that selects a model from ASVspoof 2021 according to public ranking results (i.e., EER). The call center receives a stream of calls that must be classified as real or deepfake before being transferred to the greater organization. Additionally, every detected deepfake speech sample is manually verified to check for fraudulent activity. The call centers in both scenarios receive identical streams of 4,400 [80] total monthly calls. The base-rate of deepfake calls in this environment is 1 in 1,074 (i.e., ~4 total deepfake calls in a month), a rate derived from real-world call center fraud benchmarks [81]. In SCN_A, M_{ASV-SW} is selected as this model is the top performer based on EER. In SCN_B, M_{ASV-LG} is selected, as this is the simplest model.

Examining SCN_A, M_{ASV-SW} has a near-perfect true positive rate, and correctly detects all 4 deepfake calls. However, the model incorrectly identifies 1,182 real calls as a deepfake. Calculating the Bayesian Detection Rate for this model in this specific environment gives $BDR = 0.34\%$. This means that roughly 1 in 333 calls detected as a deepfake are *actually* deepfakes. The employee tasked with manually verifying potentially fraudulent calls is overwhelmed.

Now examining SCN_B, M_{ASV-LG} results in 2 out of 4 deepfake calls being correctly detected and 38 real calls incorrectly identified as a deepfake. Calculating the BDR for this model in this specific environment gives $BDR = 0.47\%$. This suggests that 1 in 200 calls detected as a deepfake are *actually* deepfakes. As the number of false positives is much lower, the manual verification of potentially fraudulent calls is not quite as overwhelming in this scenario.

Organizations have different risk management systems and thus may prioritize individual metrics in different ways. For example, a call center may prefer a model that minimizes manual verification workload (an increase in this workload leads to a direct increase in manhours and expenses) to avoid alarm fatigue [82]. In this example, M_{ASV-LG} achieves superior performance. However, if minimizing the total number of fraudulent calls that bypass the call center is more important, M_{ASV-SW} is the better-performing model.

This scenario highlights the trade-offs often neglected when selecting detection models using a single metric. The context is key, therefore every model ***must*** be contextualized to select appropriate metrics. We observe in this example that failure to do so can translate to alarm fatigue, though negative outcomes can also include financial loss or degradation of public relations. Finally, we note that both models produce poor BDR values in this setting, highlighting the need for more robust classifiers in deepfake detection.

4.2.2 Image Deepfakes: Social Media

We theorize a secondary scenario with real-world impact. For this scenario, we first evaluated the CIFAKE baseline model; however, this model follows the trends from the previous scenario. Instead, we select the DFFD image deepfake baseline model as it is the only one with the appropriate TPR and FPR metrics reported as seen in Table 7 on our companion website. This scenario highlights the issue with single performance metrics and shows that even BDR alone is fallible. For this scenario, we use the same model at two different thresholds (i.e., TPR 90% at FPR 0.1% and TPR 83% at FPR 0.01%). We define these two models as M_1 and M_2. Each deepfake image detector is tasked with detecting fake news in social media where any deepfake image is considered fake news.

Models M_1 and M_2 are tasked with providing an end user with a prediction on whether an image is a deepfake or real. The end user may then decide to fact-check this particular piece of media. As such, fake news in this scenario is relatively inexpensive in that it has low "immediate risk" and a model falsely identifying media as a deepfake has low impact. On the other hand, a model must be able to accurately predict every piece of fake news correctly to ensure complete coverage. The base-rate of fake news on social media is unknown; however, Allcott et al. [83] find that from a list of fake news websites over half of all images proved to be fake news. Following this, the models are only tasked with predicting on media that is hosted on previously known fake news sites. As such, we will assume a high base-rate for fake news on social media as 1 in 2. We assume that each model sees the same 1,000 samples of media (i.e., 500 real samples and 500 fake news samples).

Calculating the BDR for M_1: $BDR = 99.89\%$ and the BDR for M_2: $BDR = 99.99\%$. In this scenario M_1 flags fewer samples correctly than M_2 as M_2 has a higher BDR. However, when contextualized to this scenario M_1 is the clear best performer. To highlight this, consider M_2 which detects 415 of the 1,000 samples as fake news and does not misclassify any real samples as fake news. On the other hand, M_1 detects 450 of the 1,000 samples as fake news and also does not misclassify any real samples as fake news. As neither model misclassified any real samples as fake news and M_1 detected ~10% more fake samples than M_2, the model with the lower BDR (i.e., M_1) is the best model for this scenario. Additionally, as a false positive in this scenario is considered low-risk, the choice of M_1 stands, even as the number of processed images grows and misclassifications appear.

4.3 Ethical Consideration

Serious ethical considerations arise when researching deepfake generation, datasets, and defenses; each deserving its own ethics discussion. As we analyze deepfake datasets, we contextualize the ethics of dataset generation, collection, and propagation using the Menlo Report's [84] pillars: *Respect*

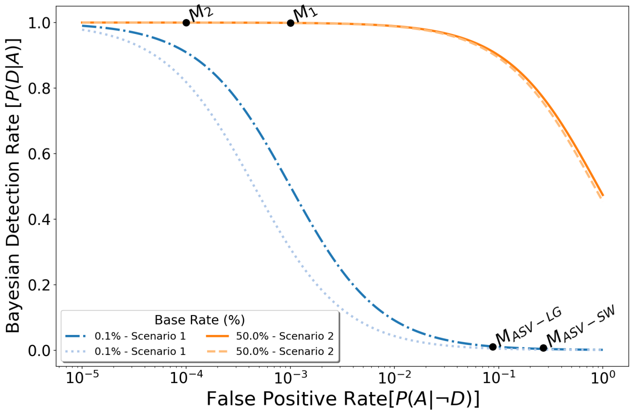


Figure 6: Scenarios from Sections 4.2.1 and 4.2.2 plotted against their base-rates and BDRs. We show that major improvements in model performance are required in the base-rate environment similar to the call-center scenario.

for Persons, *Beneficence*, *Justice*, and *Respect for Laws*.

Respect for Persons: Research involving human subjects should properly disclose the risks of the research to each participant. Whether collecting deepfakes in-the-wild or generating deepfakes from participants, researchers should make their best effort to identify and obtain consent from subjects. Further, when providing deepfake datasets for reproducibility and future research, subjects should have a clear understanding of how this data can be used.

Beneficence: Minimizing harm to subjects while maximizing the overall benefit of the research becomes increasingly difficult as various types of deepfakes can have adverse effects on the subjects of deepfakes. For example, deepfake pornography has extremely negative effects on victims [85]. Researchers have to be particularly careful in building datasets designed to encourage the development of these defenses. Publishing datasets, especially containing sensitive information, can create harmful effects for the subjects.

Justice: Each person deserves equal consideration in the course of a study. Deepfake datasets should clearly define how their populations were collected and identify inherent biases arising from this process. Researchers should consider how they can provide accurate representation of populations, especially of vulnerable and marginalized communities. Specifically, considering *who is left out* when creating a deepfake dataset allows for a more ethically sound foundation.

Respect for Laws: Including in-the-wild samples in a dataset and propagating that dataset is, at best, non-consensual and at worst, illegal [86]. Unfortunately, laws and public policy surrounding the creation and use of deepfakes has yet to be enacted at a wide scale [87]. Future policy changes and laws could greatly affect how deepfake datasets are created.

Of the forty-five deepfake datasets we surveyed, only eleven discussed ethical considerations of their dataset with none mentioning the Menlo Report. Moreover, some of these deepfake datasets train using corpora that predate deepfakes; thus, participant consent in these corpora does not include deepfake ethical considerations. While these datasets offer a foundation for deepfake research, a more thorough community discussion on the ethics of creating deepfake datasets and how the community builds upon them is necessary.

5 Recommendations for Future Datasets

5.1 Metrics

Metric Selection: Many metrics may be appropriate when applied within a specific context, but not all metrics are applicable within every context. There is *no single metric* that fully represents the performance of a model.

In general, for class-imbalanced datasets precision-recall curves and AUPRC has been shown to more accurately describe the performance of the minority class. However, the use of this metric is critical. The minority class is usually the class of import, or the class a detector is created to detect, but in the case of speech deepfake datasets, the important class is the overwhelming majority. Therefore, simply applying AUPRC to view the performance of deepfake detection in these datasets will produce non-meaningful results. The baseline for a precision-recall curve and AUPRC is the percentage of samples in the viewed class, which for example is 97% in the ASVspoof 2021 evaluation set. The difference between a perfect detector and the baseline is only 3%.

Most deepfake datasets suggest using EER and accuracy as simple performance evaluators, but both fail to honestly and completely measure model performance. While convenient as a means of providing an easy comparison, this approach largely results in meaningless results. More critically, it subsequently hinders the ability to adopt these detection models in the real world and hides the need for even better detectors.

To combat this, we recommend that future datasets suggest the use of an ensemble of metrics such as the precision-recall curve and AUPRC for the minority class (if the minority class is the important class), false positive rate, true positive rate, ROC curve/AUROC, and the BDR. Additionally, we recommend posting the raw scores output from the model, such that future researchers may calculate *any* metric for comparison. This approach ensures that model results may be contextualized to fully evaluate performance.

Metric Reporting Template: We create and publish a boilerplate template to be used for future detection mechanisms. We provide a concise example of this in Table 4 and a full example in Table 4 on our companion website shows this template as populated by the scores files from all models trained throughout Section 3.2. Using this template allows for a quick comparison of class distributions in training and test sets and

														BDR							
														Base-Rates							
	Trn Set	**Trn Distro**	**Eval Set**	**Eval Distro**	**EER**	**TPR**	**FPR**	**TNR**	**FNR**	**Prec**	**Recall**	**F1**	**AUPRC**	*0.01*	*0.1*	*1*	*10*	*25*	*50*	*75*	*90*
RawNet2	ASV	90F/10R	ASV	97F/3R	22.1	95.9	36.2	63.8	4.1	68	79.9	72.2	54.7	0	0.3	2.6	22.7	46.9	72.6	88.8	96
LFCC-LCNN	ASV	90F/10R	ASV	97F/3R	22.9	94.7	41.7	58.3	5.3	64.1	76.5	68	47.5	0	0.2	2.2	20.2	43.1	69.5	87.2	95.3
SSL-wav2vec	ASV	90F/10R	ASV	97F/3R	4.1	99.9	26.9	73.1	0.1	97.8	86.5	91.3	92.8	0	0.4	3.6	29.2	55.3	78.8	91.8	97.1
CQCC-GMM	ASV	90F/10R	ASV	97F/3R	25.4	44.2	13.1	86.9	55.8	65.6	52.3	35.7	27.8	0	0.3	3.3	27.2	52.9	77.1	91	96.8
LFCC-GMM	ASV	90F/10R	ASV	97F/3R	25.5	44.9	8.8	91.2	55.1	52.6	68.1	36.5	27.3	0.1	0.5	4.9	36.2	63	83.6	93.9	97.9

Table 4: Concise template of dataset statistic and metric reporting.

quantifies the performance as multiple base-rates. This template facilitates the mapping of any detector onto a figure similar to Figure 4, which creates a visual representation of base-rate performance.

5.2 Consider the Base-Rate

We show the importance of contextualizing the outputs of detectors within base-rates. The baseline models examined in this paper perform reasonably well within the scope of a base-rate that matches the distributions of the evaluation set but break down when tasked with inferring other distributions.

As an in-depth evaluation to measure the base-rate of deepfakes in the wild is an open research problem, there are a few potential stop-gaps for contextualizing datasets within base-rates. None of these suggestions change any requirements of datasets, as this is a post-processing step to change the viewing angle of each compared detector. As we show in Section 4.2, detectors that have substantially worse EER values may perform better in a base-rate-aware scenario. The first, and the most straightforward option, is to consider a range of base-rates that cover class imbalance in both classes similar to our experiments in Section 3.2.4. A second option is to justify the class distribution being presented in a dataset, generally in the form of a clearly defined use case. We recommend setting additional base rates at an order of magnitude above and below that of the evaluation set (i.e., for a 1% base-rate, provide three base-rates of 0.1%, 1%, and 10%).

Deepfake datasets should provide base-rate assumptions to put all results into an appropriate context. Doing so will allow for contending models to follow an explicit threat model while also providing interested organizations with the information necessary to decide if a model is appropriate for their setting. Currently, such datasets only offer extraordinarily high base-rates of deepfake samples in their datasets.

5.3 Dataset Class Imbalance

Dataset class imbalance is not a new issue in machine learning and is often a product of the difficulty of capturing samples of the minority class [88–90]. However, due to the issues highlighted in Section 4.2, deepfake datasets suggest that real samples are difficult to gather. Regardless, many techniques exist to alleviate the impact of class imbalance from simple resampling [91–93] to data augmentation [94–96] and data generation [97, 98]. Class imbalance is not inherently negative; however, leaving class imbalance unacknowledged leads to misrepresentation of results. A simple solution would be to increase the amount of human-generated media in future datasets, regardless if doing this increases the complexity of curating. This allows for a balanced dataset and a reduction of misinterpretation of model performance.

5.4 Comparing Via Contextual Importance

Comparison & Context: To reduce the chasm between detection mechanism development and real-world deployment, a more representative method of comparing model performance must be adopted. Towards this goal, strict contextualization allows for meaningful comparison of model efficacy. First, datasets should suggest an ensemble of metrics for comparators to optimize. Second, these datasets should define a use case or scenario for contextualization. Third, these metrics must then be weighted by importance. Finally, multiple base-rates should be selected for contextualization to cover any differences in measured base-rates and in-the-wild realities.

Using the call center scenario from Section 4.2.1 as an example, we showcase a simple comparison example for completeness. As the base-rate in this scenario is measured as 0.001%, three base-rates should be selected: 0.0001%, 0.001%, and 0.01%. As one possible approach, we propose a weighted ranking by setting a maximum allowed FPR and filtering models based on this threshold. The remaining models may then be ordered by decreasing TPR and/or BDR values.

Use Case: Understanding the base-rate of incidence of an attack is tantamount to clearly defining a use case. In Section 4.2, we outline scenarios with explicitly defined base-rates. Additionally, we quantify the penalty of false positives and false negatives in the form of manual reviewer fatigue and correctness. We show in these examples that traditional comparisons between models are meaningless in this context. This shows that the use case is the most important criterion to consider when designing datasets.

5.5 Using Contextless Datasets

While we suggest guidelines for the creation of future security-aware datasets, there exists the issue of the current datasets and their usefulness. Towards this, our methodology for dataset creation may be applied. First, when publishing results, make an ensemble of metrics and the raw scores file available. Second, select a wide range of base-rates to test the detection performance. If the base-rate is known, then select an order of magnitude above and below. However, if the base-rate is unknown then select a wide range of base-rates such as we do in Section 3.2.4 (e.g., (90%, 75%, 50%, 25%, 10%, 1%, 0.1%, 0.01%)). Following these steps will allow for a more meaningful comparison with future work.

6 Conclusion

Deepfake media (e.g., video/image/speech deepfakes) represents a growing threat to our trust in information. Mechanisms that facilitate community interaction and spur innovation for the detection of deepfakes are necessary. However, the confluence of data in this space makes the selection of datasets non-trivial and these datasets inadvertently promote unrealistic outputs. To understand this, we systematize the space of deepfake media and compare that to anomaly detection in the form of network intrusion detection. Through this systematization, we demonstrate the vast discrepancy between the class distribution of these two fields. Specifically, we examine the speech deepfake datasets as they are the overwhelming outliers in class distribution. We examine the most popular datasets and show that this imbalance creates bias and vastly over-represents model performance. Such bias is hidden by the use of limited metrics, specifically EER, as the sole metric for comparison and the lack of consideration of base-rates. We apply the models examined in a real-world base-rate-aware scenario to show significant performance issues in the form of false positives and alert fatigue. We thus recommend that raw model score files be published for evaluating future methods, that the context in which a model is deployed be well-defined, and that models be evaluated across varying base-rates. The failure to follow these recommendations risks entrenching inaccurate metrics used by multiple research communities, to the ultimate detriment of all detection-based research.

Acknowledgments

The authors thank our anonymous reviewers and our shepherd for their valuable comments and suggestions. This work was supported in part by the Office of Naval Research under grant number ONR-OTA N00014-21-1-2658 and the National Science Foundation under grant number NSF CNS-2206950. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research and/or the National Science Foundation.

References

[1] Catherine Stupp. Fraudsters Used AI to Mimic CEO's Voice in Unusual Crime. *Wall Street Journal*, 2019.

[2] Aengus Collins. Forged authenticity: governing deepfake risks. Technical report, EPFL International Risk Governance Center, 2019.

[3] Bobby Allyn. Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia, 2022.

[4] Pete Syme. Elon musk's past statements about self-driving safety could feasibly be deepfakes, tesla lawyers told court. https://www.businessinsider.com/tesla-lawyers-elon-musk-autopilot-safety-statements-could-be-deepfakes-2023-4, 2023.

[5] Stefan Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security*, 2000.

[6] Kristopher Kendall. *A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems*. PhD thesis, Massachusetts Institute of Technology, 1999.

[7] Prasanta Gogoi, Monowar H Bhuyan, DK Bhattacharyya, and Jugal K Kalita. Packet and Flow Based Network Intrusion Dataset. In *Contemporary Computing: 5th International Conference*, 2012.

[8] Eduardo K Viegas, Altair O Santin, and Luiz S Oliveira. Toward a Reliable Anomaly-Based Intrusion Detection in Real-World Environments. *Computer Networks*, 2017.

[9] Nour Moustafa and Jill Slay. UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set). In *2015 Military Communications and Information Systems Conference*, 2015.

[10] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *International Conference on Information Systems Security and Privacy*, 2018.

[11] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Stefanos Gritzalis. Intrusion Detection

in 802.11 Networks: Empirical Evaluation of Threats and a Public Dataset. *IEEE Communications Surveys & Tutorials*, 2016.

[12] Elaheh Biglar Beigi, Hossein Hadian Jazi, Natalia Stakhanova, and Ali A. Ghorbani. Towards Effective Feature Selection in Machine Learning-Based Botnet Detection Approaches. In *2014 IEEE Conference on Communications and Network Security*, 2014.

[13] Markus Ring, Sarah Wunderlich, Dominik Grüdl, Dieter Landes, and Andreas Hotho. Creation of Flow-Based Data Sets for Intrusion Detection. *Journal of Information Warfare*, 2017.

[14] Sebastian Garcia, Martin Grill, Jan Stiborek, and Alejandro Zunino. An Empirical Comparison of Botnet Detection Methods. *computers & security*, 2014.

[15] Mouhammd Alkasassbeh, Ghazi Al-Naymat, Ahmad BA Hassanat, and Mohammad Almseidin. Detecting Distributed Denial of Service Attacks Using Data Mining Techniques. *International Journal of Advanced Computer Science and Applications*, 2016.

[16] Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, and Ali A Ghorbani. Toward Developing a Systematic Approach to Generate Benchmark Datasets for Intrusion Detection. *computers & security*, 2012.

[17] Jungsuk Song, Hiroki Takakura, Yasuo Okabe, Masashi Eto, Daisuke Inoue, and Koji Nakao. Statistical Analysis of Honeypot Data and Building of Kyoto 2006+ Dataset for NIDS Evaluation. In *Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, 2011.

[18] Raman Singh, Harish Kumar, and RK Singla. A Reference Dataset for Network Traffic Activity Based Intrusion Detection System. *International Journal of Computers Communications & Control*, 2015.

[19] Rohini Sharma, RK Singla, and Ajay Guleria. A New Labeled Flow-Based DNS Dataset for Anomaly Detection: PUF Dataset. *Procedia computer science*, 2018.

[20] Sangeeta Bhattacharya and S Selvakumar. SSENet-2014 Dataset: A Dataset for Detection of Multiconnection Attacks. In *Eco-friendly Computing and Communication Systems*, 2014.

[21] Markus Ring, Sarah Wunderlich, Deniz Scheuring, Dieter Landes, and Andreas Hotho. A Survey of Network-Based Intrusion Detection Data Sets. *Computers & Security*, 2019.

[22] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *IEEE International Workshop on Information Forensics and Security*, 2018.

[23] Pavel Korshunov and Sebastien Marcel. DeepFakes: a New Threat to Face Recognition? Assessment and Detection, 2018.

[24] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *International Conference on Digital Image Computing: Techniques and Applications*, 2022.

[25] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[26] Jiajun Huang, Xueyu Wang, Bo Du, Pei Du, and Chang Xu. DeepFake MNIST+: A DeepFake Facial Animation Dataset. In *International Conference on Computer Vision Workshops*, 2021.

[27] Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Deephy: On deepfake phylogeny. In *International Joint Conference on Biometrics*, 2022.

[28] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset, 2020.

[29] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *International Conference on Computer Vision*, 2019.

[30] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[31] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM international conference on multimedia*, 2020.

[32] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint*, 2021.

[33] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the Detection of Digital Face

Manipulation. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[34] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[35] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. KoDF: A Large-scale Korean DeepFake Detection Dataset. In *International Conference on Computer Vision*, 2021.

[36] Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. ArtiFact: A Large-Scale Dataset with Artificial and Factual Images for Generalizable and Robust Synthetic Image Detection. *arXiv preprint*, 2023.

[37] Jordan J Bird and Ahmad Lotfi. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *arXiv preprint*, 2023.

[38] Kai Nakamura, Sharon Levy, and William Yang Wang. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Language Resources and Evaluation Conference*, 2020.

[39] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces. In *International Joint Conference on Artificial Intelligence*, 2020.

[40] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *Computer Vision and Pattern Recognition Workshops*, 2017.

[41] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *International Conference on Acoustics, Speech and Signal Processing*, 2019.

[42] Tomi Kinnunen, Md Sahidullah, Hector Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In *Interspeech*, 2017.

[43] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint*, 2019.

[44] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint*, 2021.

[45] Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Hong-Goo Kang, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. Sasv challenge 2022: A spoofing aware speaker verification challenge evaluation plan. *arXiv preprint*, 2022.

[46] Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint*, 2021.

[47] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint*, 2022.

[48] Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan, Jianhua Tao, Tao Wang, Shiming Wang, Le Xu, and Ruibo Fu. Fad: A chinese dataset for fake audio detection. *arXiv preprint*, 2022.

[49] Roland Baumann, Khalid Mahmood Malik, Ali Javed, Andersen Ball, Brandon Kujawa, and Hafiz Malik. Voice spoofing detection corpus for single and multi-order audio replays. *Computer Speech & Language*, 2021.

[50] Yuan Gong, Jian Yang, Jacob Huber, Mitchell MacKnight, and Christian Poellabauer. Remasc: realistic replay attack corpus for voice controlled systems. *arXiv preprint*, 2019.

[51] Pavel Korshunov, Sébastien Marcel, Hannah Muckenhirn, André R Gonçalves, AG Souza Mello, RP Velloso Violato, Flávio O Simoes, M Uliani Neto, Marcus de Assis Angeloni, José Augusto Stuchi, et al. Overview of btas 2016 speaker anti-spoofing competition. In *IEEE international conference on biometrics theory, applications and systems*, 2016.

[52] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. *Advances in neural information processing systems*, 2018.

[53] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Conference of the international speech communication association*, 2015.

[54] Ricardo Reimao and Vassilios Tzerpos. For: A dataset for synthetic speech detection. In *International Conference on Speech Technology and Human-Computer Dialogue*, 2019.

[55] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi. The voice conversion challenge 2016. In *Interspeech*, 2016.

[56] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint*, 2018.

[57] Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint*, 2020.

[58] Dora M Ballesteros, Yohanna Rodriguez, and Diego Renza. A dataset of histograms of original and fake voice recordings (h-voice). *Data in brief*, 2020.

[59] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. Add 2022: the first audio deep synthesis detection challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[60] Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, et al. Add 2023: the second audio deepfake detection challenge. *arXiv preprint*, 2023.

[61] Jiangyan Yi, Ye Bai, Jianhua Tao, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu. Half-truth: A partially fake audio detection dataset. *arXiv preprint*, 2021.

[62] Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. Fmfcc-a: a challenging mandarin dataset for synthetic speech detection. In *International Workshop on Digital Watermarking*, 2021.

[63] Tomi Kinnunen, Md Sahidullah, Mauro Falcone, Luca Costantini, Rosa González Hautamäki, Dennis Thomsen, Achintya Sarkar, Zheng-Hua Tan, Héctor Delgado, Massimiliano Todisco, et al. Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. In *IEEE International conference on acoustics, speech and signal processing*, 2017.

[64] Richard P Lippmann, David J Fried, Isaac Graf, Joshua W Haines, Kristopher R Kendall, David McClung, Dan Weber, Seth E Webster, Dan Wyschogrod, Robert K Cunningham, et al. Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation. In *DARPA Information Survivability Conference and Exposition*, 2000.

[65] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 2020.

[66] Shridatt Sugrim, Can Liu, Meghan McLean, and Janne Lindqvist. Robust performance metrics for authentication systems. In *Network and Distributed Systems Security Symposium*, 2019.

[67] Daniel Olszewski, Allison Lu, Carson Stillman, Kevin Warren, Cole Kitroser, Alejandro Pascual, Divyajyoti Ukirde, Kevin Butler, and Patrick Traynor. "Get In Researchers; We're Measuring Reproducibility": A Reproducibility Study of Machine Learning Papers in Tier 1 Security Confernces. 2023.

[68] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jeeweon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint*, 2022.

[69] Héctor Delgado, Nicholas Evans, Jeeweon Jung, Tomi Kinnunen, Ivan Kukanov, Kong Lee, Xuechen Liu, Hyejin Shim, Md Sahidullah, and Hemlata Tak. Asvspoof 2021 baseline cm and evaluation package. `https://github.com/asvspoof-challenge/2021`, 2021.

[70] Hemlata Tak. Ssl anti-spoofing using wav2vec 2.0 and data augmentation. `https://github.com/TakHemlata/SSL_Anti-spoofing`, 2022.

[71] Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan, Jianhua Tao, Tao Wang, Shiming Wang, Le Xu, and Ruibo Fu. Cfad baseline systems. `https://github.com/ADDchallenge/CFAD`, 2022.

[72] Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, et al. Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *arXiv preprint*, 2021.

[73] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. *arXiv preprint*, 2020.

[74] Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, Lorenzo Cavallaro, et al. Tesseract: Eliminating experimental bias in malware classification across space and time. In *USENIX Security Symposium*, 2019.

[75] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *IEEE international conference on acoustics, speech and signal processing*, 2015.

[76] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP*, 2022.

[77] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 1993.

[78] Keith Ito. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[79] KYUBYONG PARK. The world english bible. https://www.kaggle.com/datasets/bryanpark/the-world-english-bible-speech-dataset, 2018.

[80] Tony Zerucha. Fraud in call centers changing but still on the rise in 2022: Pindrop. https://www.crowdfundinsider.com/2022/02/187157-fraud-in-call-centers-changing-but-still-on-the-rise-in-2022-pindrop/, 2022.

[81] 10 call center benchmarks. https://www.liveagent.com/research/call-center-benchmarks/, 2023.

[82] Judy Edworthy and Elizabeth Hellier. Alarms and human behaviour: implications for medical alarms. *British Journal of Anaesthesia*, 2006.

[83] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 2017.

[84] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The menlo report. *IEEE Security & Privacy*, 10(2):71–75, 2012.

[85] Sophie Maddocks. 'a deepfake porn plot intended to silence me': exploring continuities between pornographic and 'political'deep fakes. *Porn Studies*, 7(4):415–423, 2020.

[86] Edvinas Meskys, Julija Kalpokiene, Paul Jurcys, and Aidas Liaudanskas. Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, 2020.

[87] The high stakes of deepfakes: The growing necessity of federal legislation to regulate this rapidly evolving technology, 2023.

[88] D Ramyachitra and Parasuraman Manikandan. Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research*, 2014.

[89] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*, 2006.

[90] Ranjit Panigrahi and Samarjeet Borah. A detailed analysis of cicids2017 dataset for designing intrusion detection systems. *International Journal of Engineering & Technology*, 2018.

[91] Razan Abdulhammed, Miad Faezipour, Abdelshakour Abuzneid, and Arafat AbuMallouh. Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic. *IEEE Sensors Letters*, 2019.

[92] Sikha Bagui and Kunqi Li. Resampling imbalanced data for network intrusion detection datasets. *Journal of Big Data*, 2021.

[93] David A Cieslak, Nitesh V Chawla, and Aaron Striegel. Combating imbalance in network intrusion datasets. In *GrC*, 2006.

[94] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 2019.

[95] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 2017.

[96] Shengyun Wei, Shun Zou, Feifan Liao, et al. A comparison on data augmentation methods based on deep learning for audio classification. In *Journal of Physics: Conference Series*, 2020.

[97] Gozde Karatas, Onder Demir, and Ozgur Koray Sahingoz. Increasing the performance of machine learning-based idss on an imbalanced and up-to-date dataset. *IEEE Access*, 2020.

[98] Hongpo Zhang, Lulu Huang, Chase Q Wu, and Zhanbo Li. An effective convolutional neural network based on smote and gaussian mixture model for intrusion detection in imbalanced dataset. *Computer Networks*, 2020.

7 Appendix

For a detailed analysis of additional material, please see our companion website at: https://sites.google.com/view/thegoodthebadandtheunbalanced.