Fake or Real

a 

b 

c 

d 

Fake or Real

a **b**

c **d**

a , c	Fake (VC)
b	Fake (TTS)
d	Real

Strengthening AI Models for Spoofed Audio Detection: An Interdisciplinary Approach Incorporating Linguistic Knowledge

Zahra Khanjani, Chloe Evered, Christine Mallinson, Vandana P. Janeja

University of Maryland, Baltimore County

ADSA 2024

Funded by NSF awards #2210011 and #2346473

Motivation

- Recent incidents of Fraud using Audio Deepfake (a type of spoofed audio)

TWO REAL-WORLD SCENARIOS

OZY MEDIA FRAUD SCHEME

- Ozy Media arranged a call with Goldman Sachs and a YouTube partner to secure a $40 million investment.
- During the call, the voice representing Ozy Media and YouTube sounded digitally altered, attempting to impersonate an executive.

https://tinyurl.com/37h3pm58

BANK ROBBERY IN HONG KONG

- A bank manager in Hong Kong authorized a $35 million transfer based on an AI-generated voice that impersonated the company director.
- The fake voice was successful in its deception, resulting in the fraudulent transfer of funds.

https://tinyurl.com/37h3pm58

SOCIETAL IMPACTS

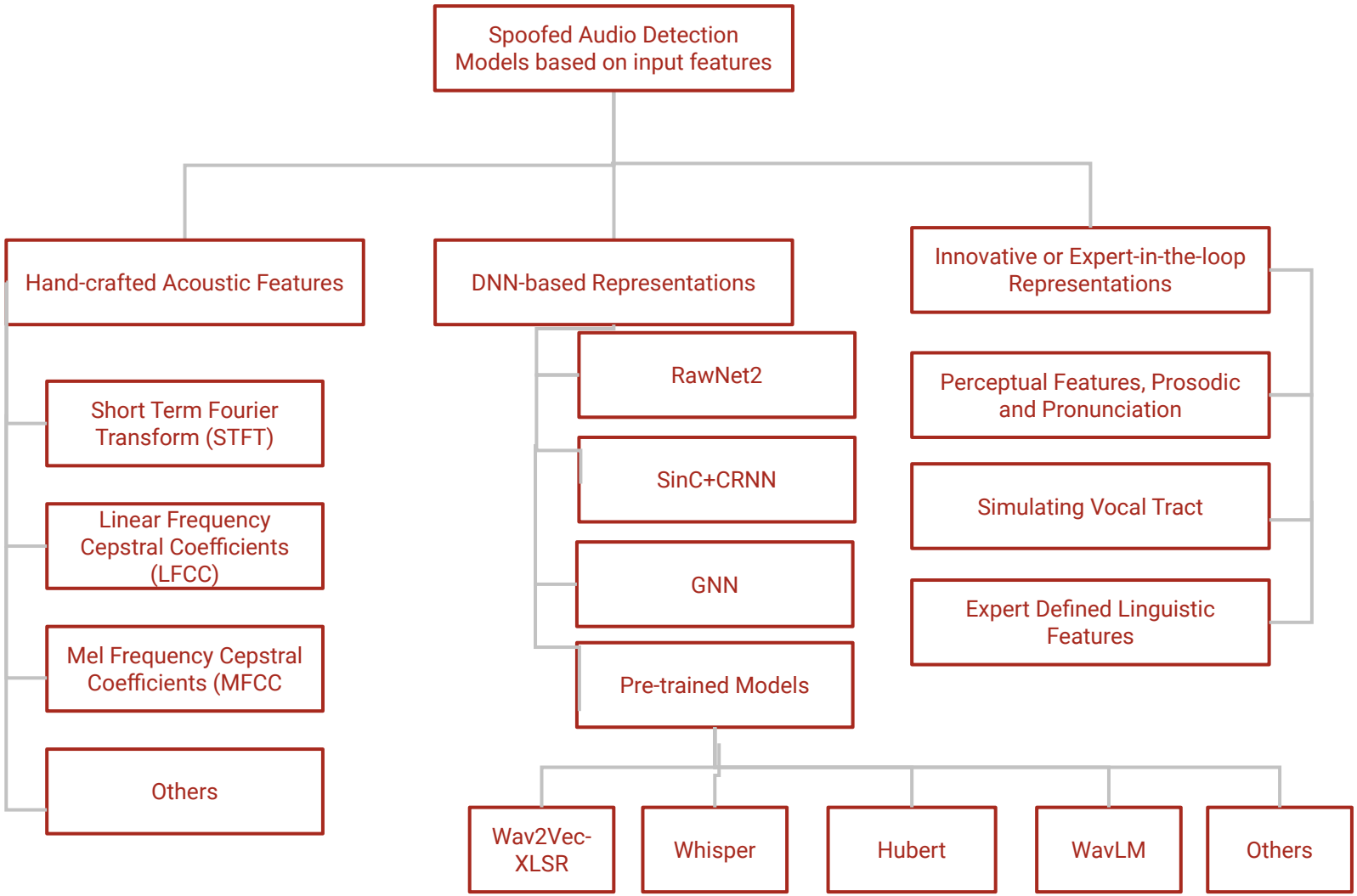
- Spoofed audio can contribute to deception and disinformation in society.
- It can undermine trust in communication channels and institutions.
- It can lead to financial losses, societal threats, fraud, impersonation, and damage to reputations.

Scammed by a video call including video and audio deepfakes impersonating his friend. He lost 4.3 million (2023)
A CEO of a U.K energy based firm lost 220,000 Euros since he thought he was on the phone with one of the executives (2019)

Motivation

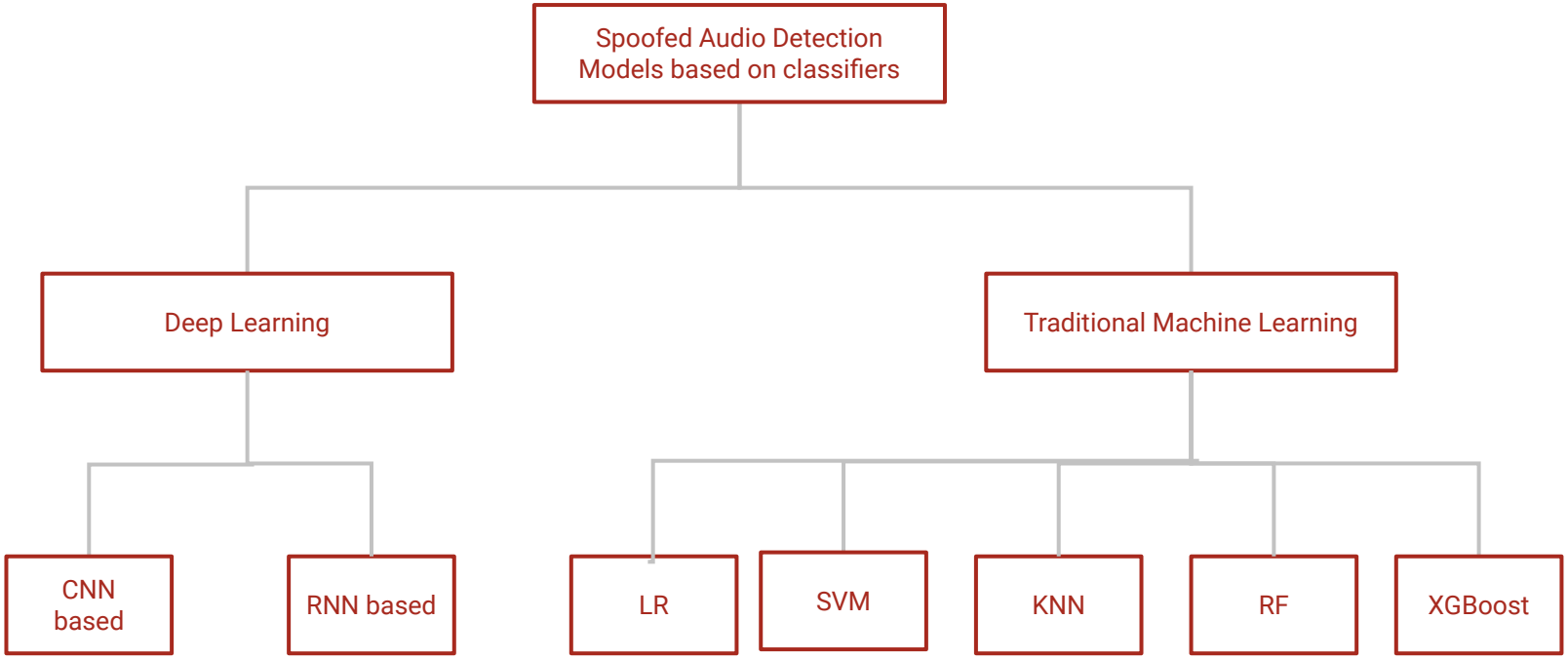
Image: Tero Vesalainen (Shutterstock)

Spoofed Audio Types (AI generated and non-AI generated)


Spoofed Audio Detection Models based on input features
Hand-crafted Acoustic Features
DNN-based Representations
Innovative or Expert-in-the-loop Representations
Short Term Fourier Transform (STFT)
Linear Frequency Cepstral Coefficients (LFCC)
Mel Frequency Cepstral Coefficients (MFCC
Others
RawNet2
SinC+CRNN
GNN
Pre-trained Models
Perceptual Features, Prosodic and Pronunciation
Simulating Vocal Tract
Expert Defined Linguistic Features
Wav2Vec-XLSR
Whisper
Hubert
WavLM
Others

Related work

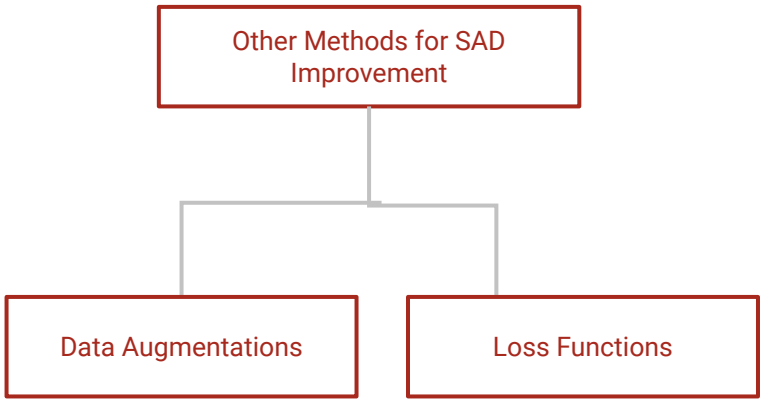
Features - Classifiers - Others



Borrelli et al., Khochare et al.

Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*, *15*(5), 155

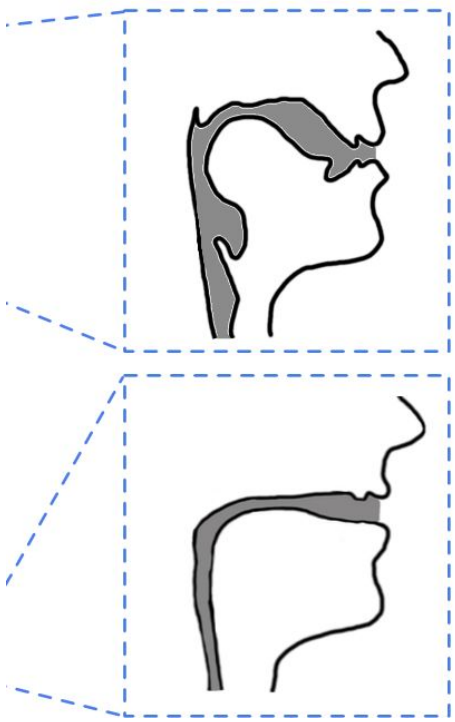
Features - Classifiers - Others



Related work

Articulatory phonetic techniques

- to identify spoofed English audio by discerning that the clips in question were impossible or highly unlikely to have been produced in a human vocal tract.
- The drawback using not only one type of attack (TTS), but also a single generative algorithm
- The figure shows An anatomical approximation of a deepfaked model (bottom), which no longer represents a regular human vocal tract (top) and instead is approximately the dimensions of a drinking straw.



Requires specialized resonator pre-processing, vowels only, training, and methodology to do the analysis and an authentic audio sample for comparison.

BLUE, L., WARREN, K., ABDULLAH, H., GIBSON, C., VARGAS, L., O'DELL, J., BUTLER, K., AND TRAYNOR, P. Who are you (i really wanna know)? detecting audio {DeepFakes} through vocal tract reconstruction. In 31st USENIX Security Symposium (USENIX Security 22) (2022), pp. 2691–2708

A place for Expert-in-the-loop

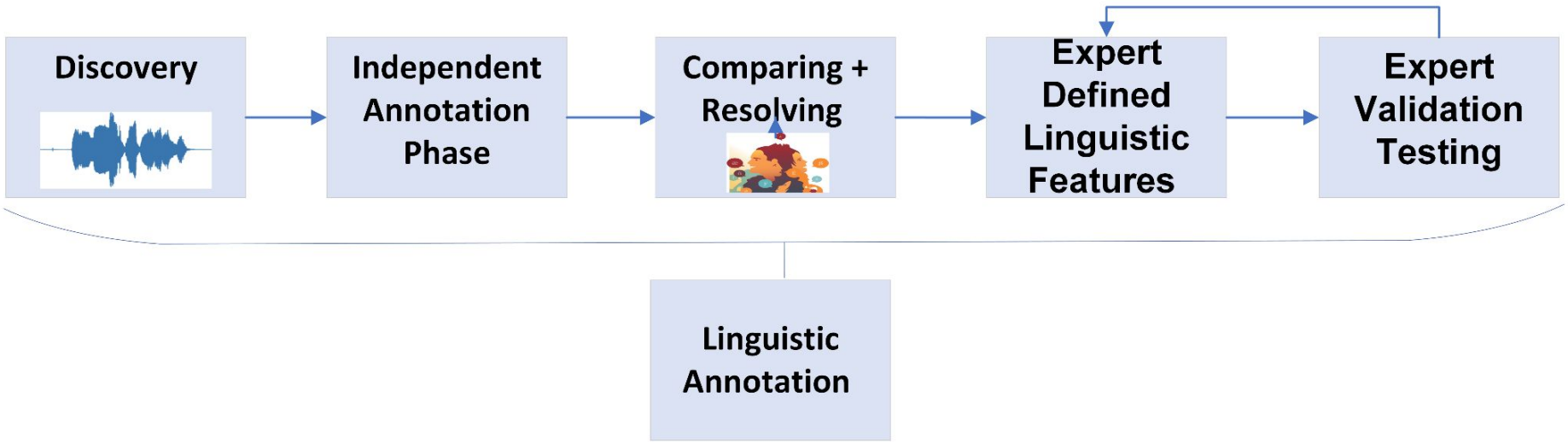
Based on the literature, two issues in Spoofed Audio Detection (SAD) is:

- Generalization (Pham et al., 2024), performance of the-state-of-the-art models drop simply by adding noise to the datasets (AlAli et al., 2023)
- Lack of multidisciplinary approaches (Boumber et al., 2024)

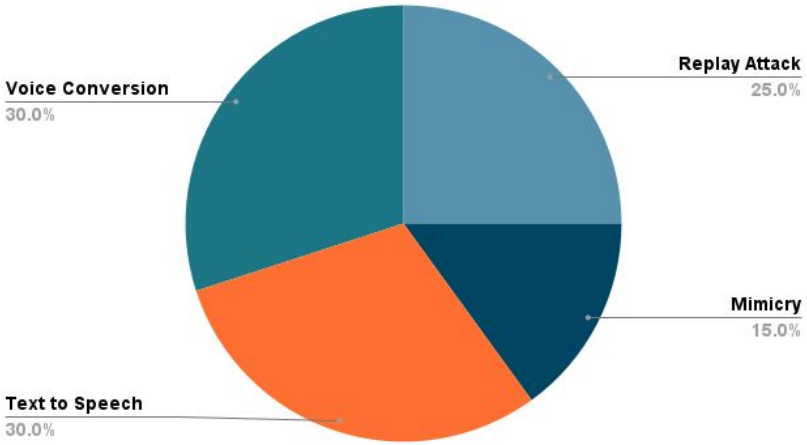
Linguistic Audio Representations based on the knowledge of sociolinguistics experts--Strengthening AI with human knowledge and Strengthening human discernment with human knowledge - our team

Expert Defined Linguistic Features (EDLFs)

Linguistic Features Annotation Methodology

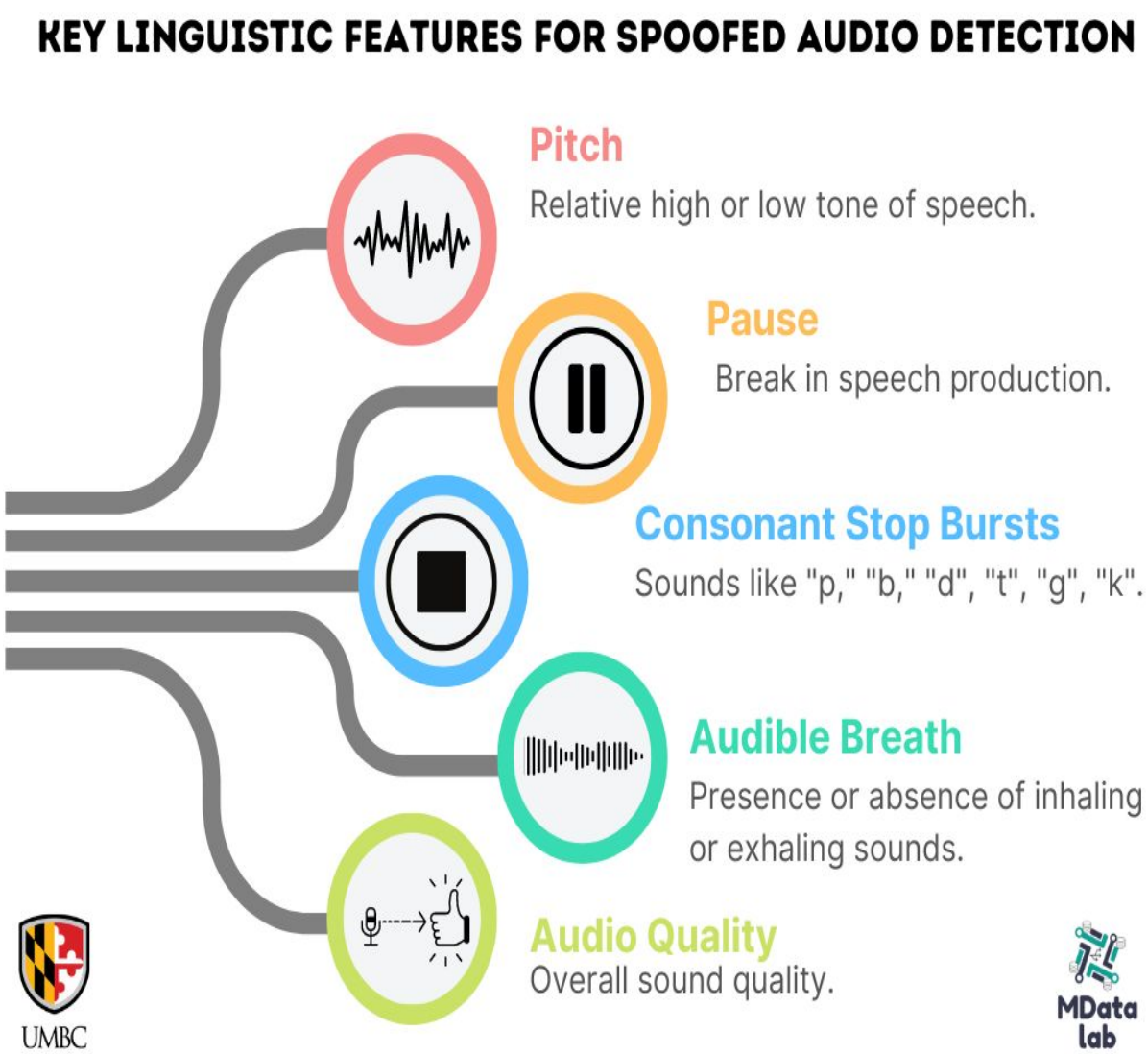


Dataset



- Multiple types of spoofed audio
- State-of-the-art VC methods included
- Subset of available datasets with added samples

A place for Expert-in-the-loop



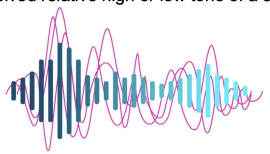
Linguistic Data Augmentation based on the knowledge of sociolinguistics experts--Strengthening AI with human knowledge and Strengthening human discernment with human knowledge

Expert Defined Linguistic Features (EDLFs)

Pitch

Pitch: the perceived relative high or low tone of a speech sample.

PITCH



- Defined for this study as the perceived relative high or low tone of the speech sample.
- Anomalous occurence of pitch-the sample received an annotation of 1
  - unusually higher or lower than expected, or
  - unusually fluctuating or inconsistent
- Normal occurrence usual or within a normal range of English language variation
  - Annotated with a 0

Pause: a break in speech production within a speech sample.

Pause

- A break in speech production within a sample.
- Anomalous Pause-the sample received an annotation of 1
  - lack of a pause where one would be expected,
  - addition of a pause where one would not be expected (such as between words of a phrase),
- Pause as usual or within a normal range of English language variation
  - annotated with a 0.

Bursts: Word-initial or word-final consonant stops

- The sounds /p/, /b/, /t/, /d/, /k/, and /g/
- Anomalous received an annotation of 1
  - lack of a burst of air where one would be expected,
  - The addition of a burst of air where one would not be expected,
  - An unusually exaggerated or truncated burst
- Production of consonant sounds perceived as usual or within a normal range of English language variation
- Annotated with a 0

Expert Defined Linguistic Features (EDLFs)

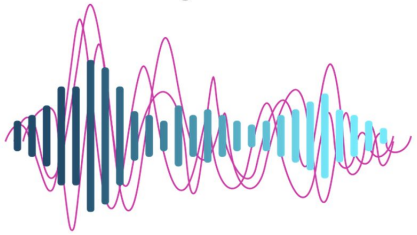
Audio Quality

Any disturbance or distortion to the speech signal

Tinny
Nasal
Echo
Compressed
Buzzing
Robotic

Pitch

Pitch: the perceived relative high or low tone of a speech sample.



Pause

Pause: a break in speech production within a speech sample.

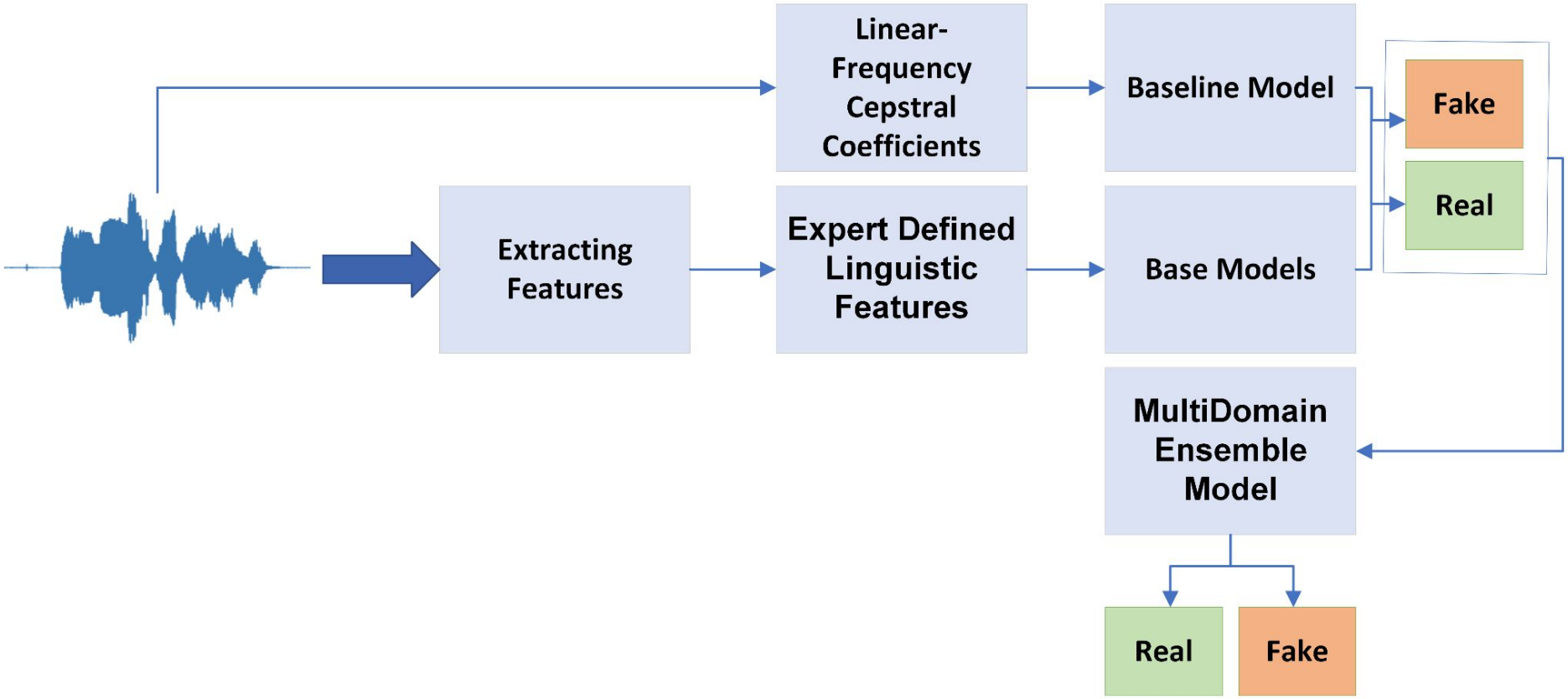
Initial and Final Consonant Bursts

Lack of a burst of air where one would be expected, the addition of a burst of air where one would not be expected, or an unusually produced burst at the beginning or end of a word.

Breath

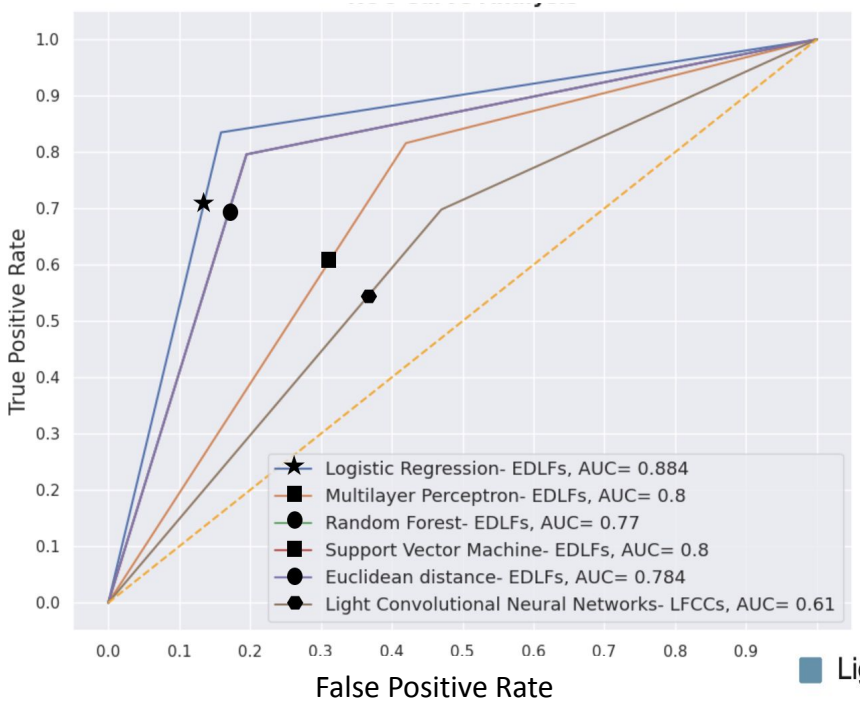
Any intake or outtake of breath

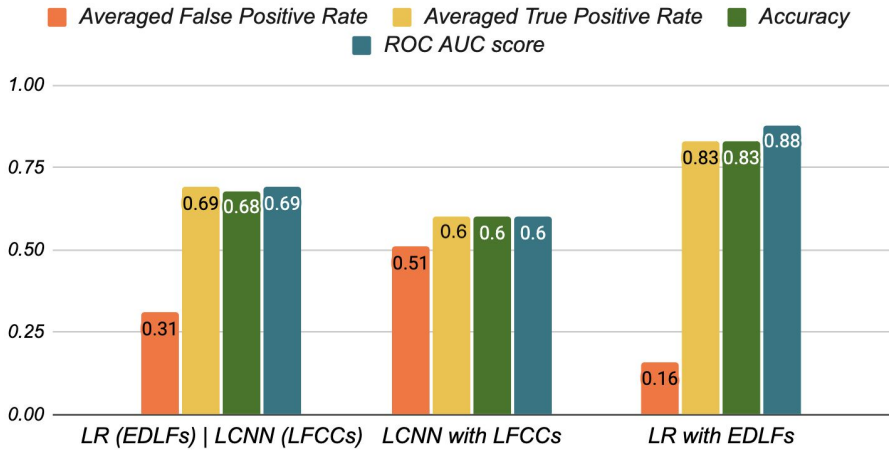
Augmenting AI models

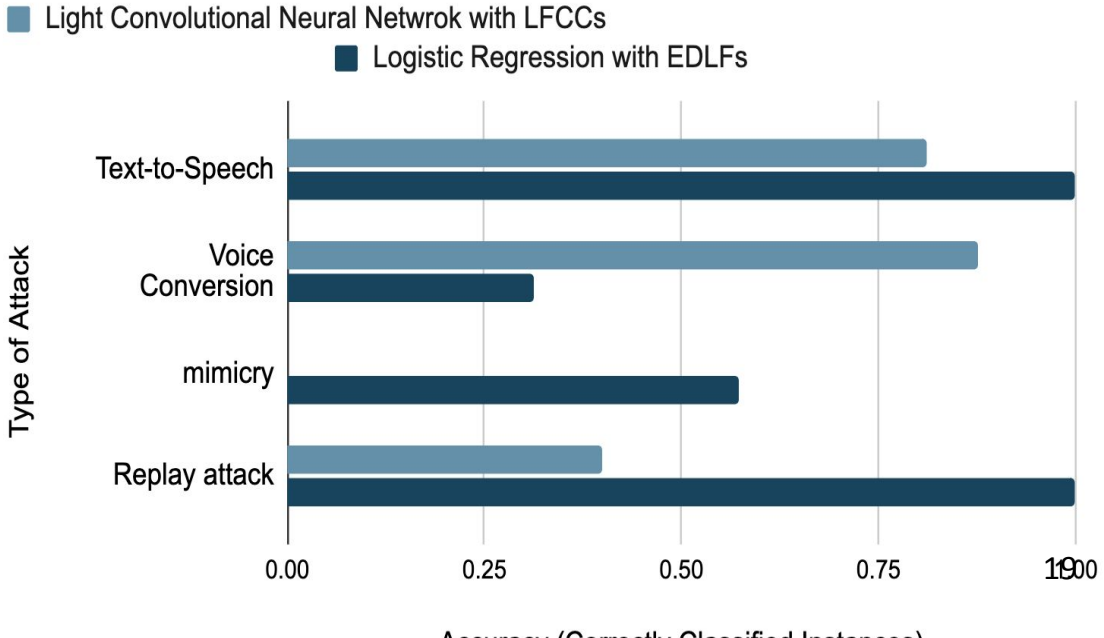


Khanjani, Z., Davis, L., Tuz, A., Nwosu, K., Mallinson, C., & Janeja, V. P. (2023, October). Learning to listen and listening to learn: Spoofed audio detection through linguistic data augmentation. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 01-06). IEEE.

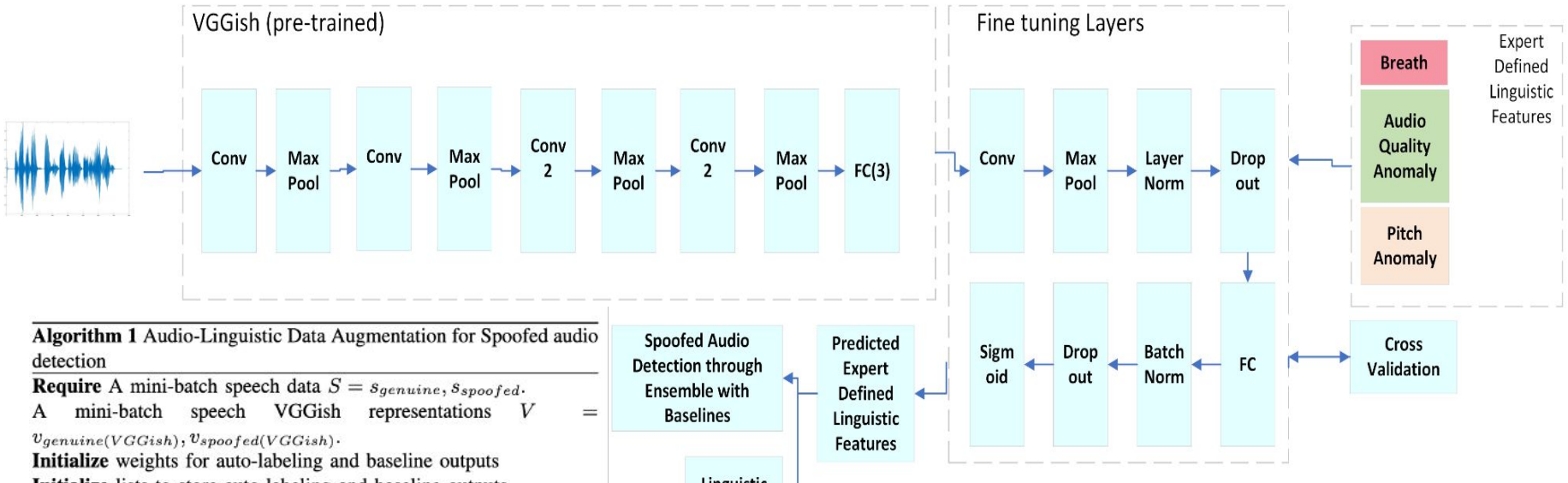
Results - Manual Labeling







Auto Labelling



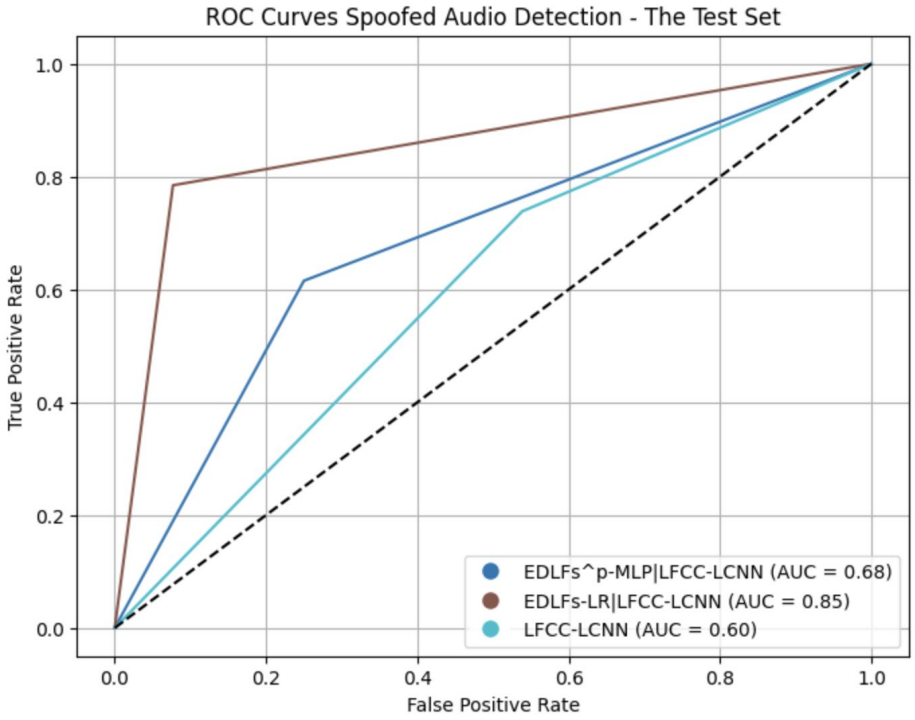
Algorithm 1 Audio-Linguistic Data Augmentation for Spoofed audio detection

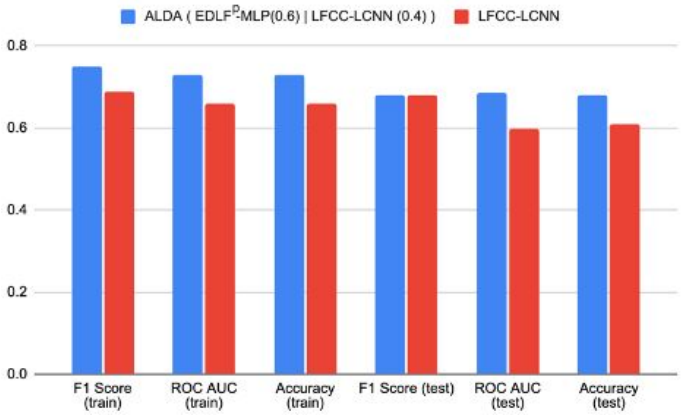
Require A mini-batch speech data $S = s_{genuine}, s_{spoofed}$.
A mini-batch speech VGGish representations $V = v_{genuine(VGGish)}, v_{spoofed(VGGish)}$.
Initialize weights for auto-labeling and baseline outputs
Initialize lists to store auto-labeling and baseline outputs
If $EDLF^p = presence - of - breath$ or $pitch - anomaly$:
$V = SMOTE(V)$,
for number of training iterations **do**
for k-th mini-batch **do**
for each audio_sample_ in V **do**
Pass through EDLF-prediction
return $EDLF^p$
Obtain baseline output:
baseline_output = GetBaselineOutput($(s_i \in S)$)
baseline_outputs.append(baseline_output)
Spoof Detection with $EDLF^p$**:**
SpoofDet_output = MLP(EDLF^p)
SpoofDet_outputs.append(SpoofDet_output)
Ensemble of the outputs:
classification_result = [weight_ALDAS ×(SpoofDet_outputs)|weight_baseline × (baseline_outputs)]
return predicted labels(spoofed, genuine)

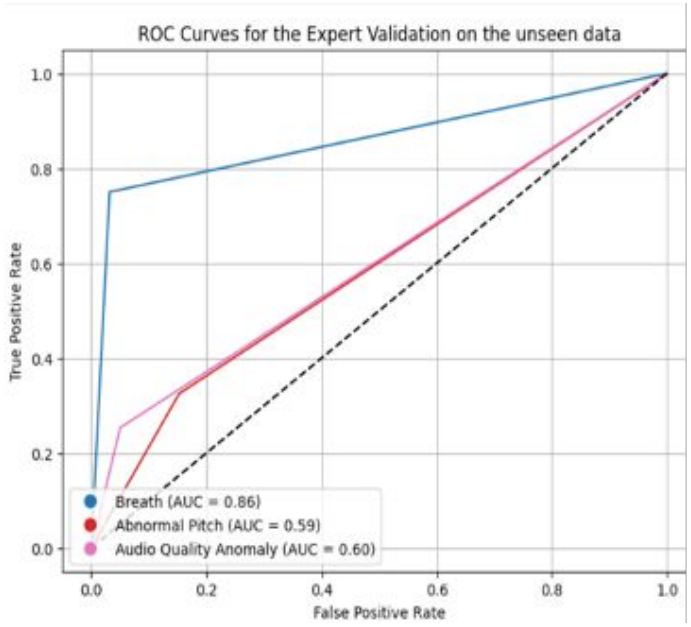
ALiRAS: Audio Linguistic Representation for Anti-Spoofing

Khanjani, Z., Mallinson, C., Foulds, J & Janeja, V. P. (2024). ALDAS: Audio-Linguistic Data Augmentation for Spoofed Audio Detection. arXiv preprint arXiv:2410.15577

Results - Auto Labeling

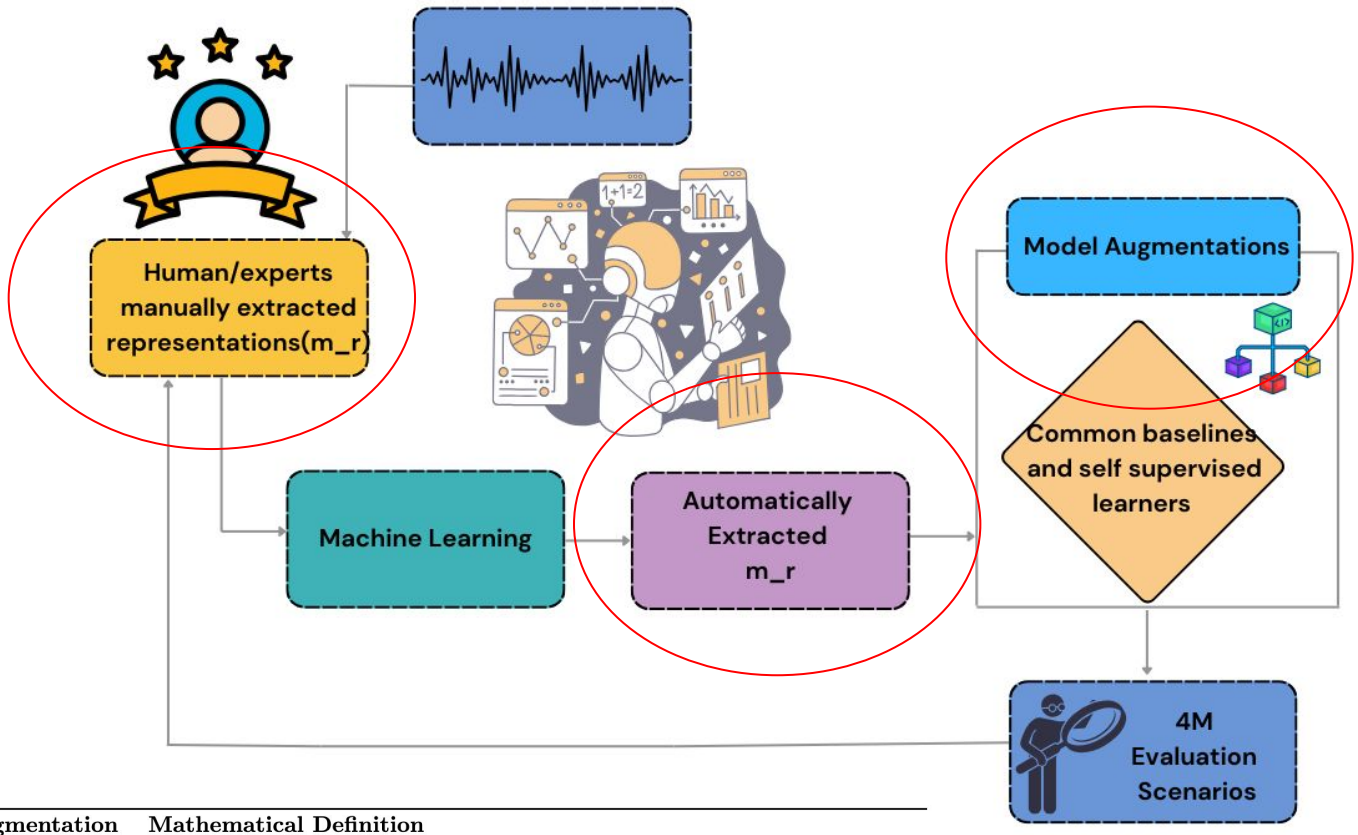






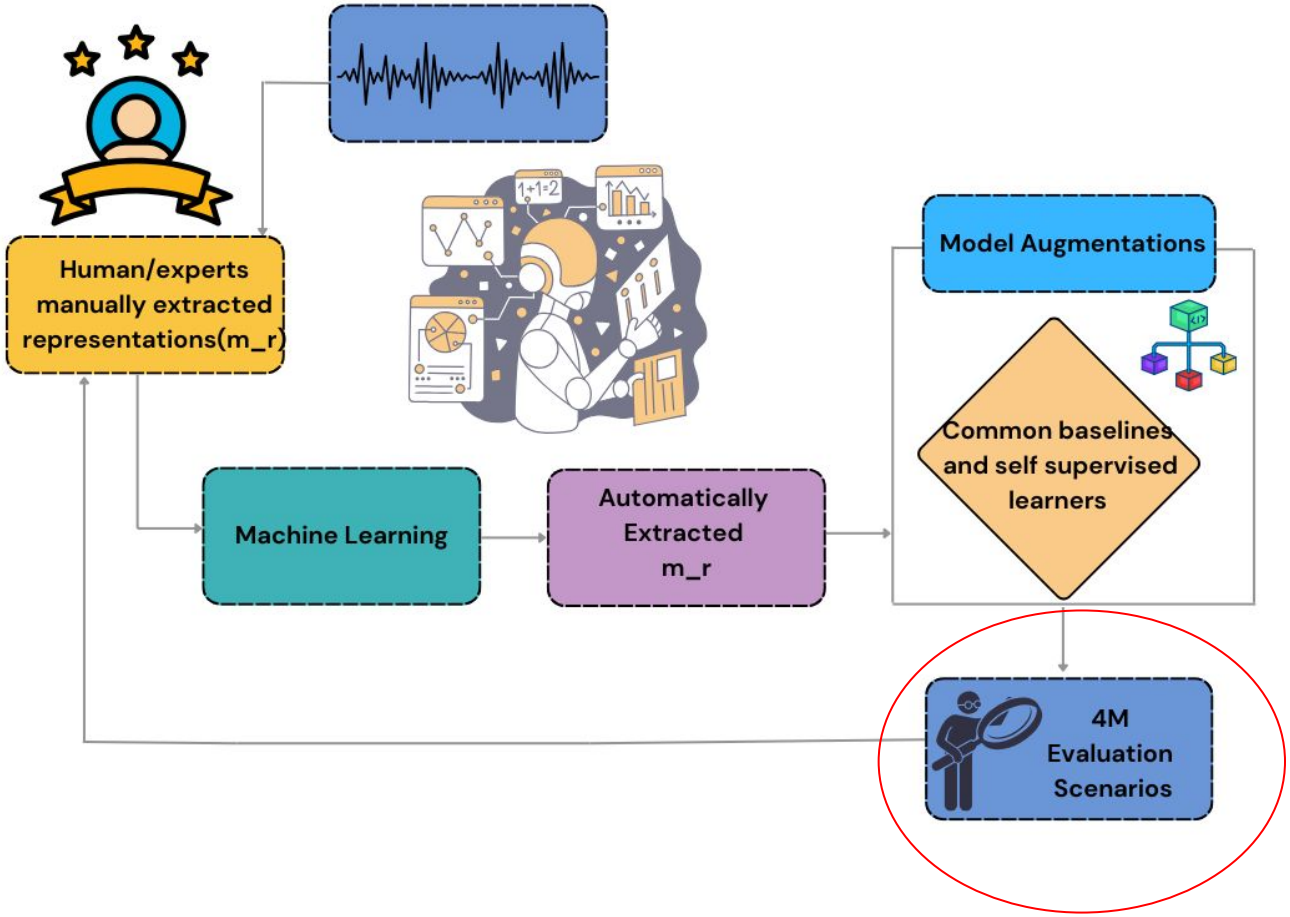
EER FOR THE BEST BASELINE ALONE, AND WHEN TRUE EDLFS AND PREDICTED EDLFS ARE INVOLVED; FOR THE TEST SET

Dataset	LFCC-LCNN	*EDLF-LR*|LFCC-LCNN	*EDLF^p-MLP*|LFCC-LCNN
Train	0.33	0.145	0.25
Test	0.39	0.14	0.31



Model Augmentation	Mathematical Definition
Ensemble Modeling	If $y_i^{pred} = spoofed \vee y_i^{pred-m-r} = spoofed$ then $\hat{y}_i = spoofed$.
Modifying Loss Function	Given the set of pairs: $(x_i = (x_{i,1}, .., x_{i,d}), y_i)_{i=1}^n$, $f : \mathbb{R}^d \to \mathbb{R}$, Such that the residual $(\|\|f(X) - y\|\|)^2$ is minimized, where $X \in \mathbb{R}^{n \times d}$ has entries $X_{i,j} = x_{i,j}$, we assume $f(X) = (f(x_i))_{i=1}^n \in \mathbb{R}^n$. $f(X) = y_{sr}$. Then, $L_{\text{modified}} = \|\|f(X) - y\|\|^2 + \lambda\|\|f(X) - y_{sr}\|\|^2$
Feature Concatenation	Given the output of the last layer k-th as: f(x) = $\text{Layer}_k(\text{Layer}_{k-1}(\ldots(\mathbf{x}))), and$ m-r, $\mathbf{z} = \begin{bmatrix}\mathbf{f(x)} \\ \text{m-r}\end{bmatrix}; then, \hat{y}_i = \text{Classifier}(z_i)$

Our Expert-in-the-loop Representation Learning and the Next Steps …



We acknowledge the contribution of Pragya Pandit for designing the infographic

The other part of the project and the team members

Can these EDLFs help human to discern Spoofed Audio better?

Faculty and Researchers
Vandana Janeja - UMBC
Christine Mallinson
Sanjay Purushotham
Anita Komlodi

PhD Students
Sara Khanjani
Noshaba Nasir Bhalli
Alabi Jamiu Ahmed
Lavon Davis

Master's Student
Lavanya Neelakandan
Pragya Pandit
Chloe Evered (Georgetown)

Undergraduate Students
Ashraf Kawooya
Kiffy Nwosu
Kavin Manivannan
Nehal Naqvi
Whitney Fils-Aime (Georgetown REU)
Gabriella Watson (Graduated)

High School Students
Jackson Means (Mt. Hebron)
Tai Akinlosotu (Mt. Hebron)

NSF

CISAAD

NSF Awards CIRC #2346473 and SaTC #2210011

Community Infrastructure to Strengthen AI for Audio Deepfake analysis

Share your thoughts about a Community Infrastructure on English Audio Deepfake research (Short survey)

Let's Connect on LinkedIn! 📱

linkedin.com/in/zahra-khanjani-data-scientist/

References

1. image from: https://www.foxbusiness.com/markets/goldman-sachs-ceo-layoffs-coming-january-report

2. BREWSTER, T. Fraudsters cloned company director's voice in $35 million bank heist, police find. Forbes, Editor's Pick 14 (2021). https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=2d60538a7559, [Online; accessed January 10, 2023].

3. https://gizmodo.com/deepfake-ai-scammer-money-wiring-china-1850461160

4. KHANJANI, Z., JANEJA, V., AND TUZ, A. Ieee-isi2023deepfake-detection. https://github.com/MultiDataLab/IEEE-ISI2023DeepFake-Detection, 2023.

5. KHANJANI, Z., WATSON, G., AND JANEJA, V. P. Audio deepfakes: A survey. Frontiers in Big Data 5 (2022)

6. WANG, R., JUEFEI-XU, F., HUANG, Y., GUO, Q., XIE, X., MA, L.,AND LIU, Y. DeepSonar: Towards effective and robust detection of AI-synthesized fake voices. In Proceedings of the 28th ACM International Conference on Multimedia (2020), MM '20, Association for Computing Machinery, pp. 1207–1216.

7. BLUE, L., WARREN, K., ABDULLAH, H., GIBSON, C., VARGAS,L., O'DELL, J., BUTLER, K., AND TRAYNOR, P. Who are you (i really wanna know)? detecting audio {DeepFakes} through vocal tract reconstruction. In 31st USENIX Security Symposium (USENIX Security22) (2022), pp. 2691–2708

8.Pham, L., Lam, P., Nguyen, T., Tang, H., Nguyen, H.,Schindler, A., Vu, C. (2024). A Comprehensive Survey with Critical Analysis for Deepfake Speech Detection. arXiv preprint arXiv:2409.15180

9. AlAli, A., & Theodorakopoulos, G. (2023, December). An RFP dataset for Real, Fake, and Partially fake audio detection. In International Conference on Cyber Security, Privacy in Communication Networks (pp. 1-15). Singapore: Springer Nature Singapore.

10. Boumber, D., Verma, R. M., Qachfar, F. Z. (2024). Blue Sky: Multilingual, Multimodal Domain Independent Deception Detection. In Proceedings of the 2024 SIAM International Conference on Data Mining (SDM) (pp. 396-399). Society for Industrial and Applied Mathematics.

11. Stupp, C. "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case" https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (accessed Sep. 12, 2024).

12. Nikki Main. Man Scammed by Deepfake Video and Audio Imitating His Friend. Web-site Name, 2023. accessed August 14, 2023, from https://gizmodo.com/deepfake-ai-scammer-money-wiring-china-1850461160.

13. Khanjani, Z., Mallinson, C., Foulds, J & Janeja, V. P. (2024). ALDAS: Audio-Linguistic Data Augmentation for Spoofed Audio Detection. arXiv preprint arXiv:2410.15577

14. Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*, *15*(5), 155

Thank You & Questions?

Funded by NSF award #2210011