Audio Spoof Detection for Indian setting

W/ DR GOKUL S KRISHNAN, DR GEETHA RAJU AND PROF B. RAVINDRAN

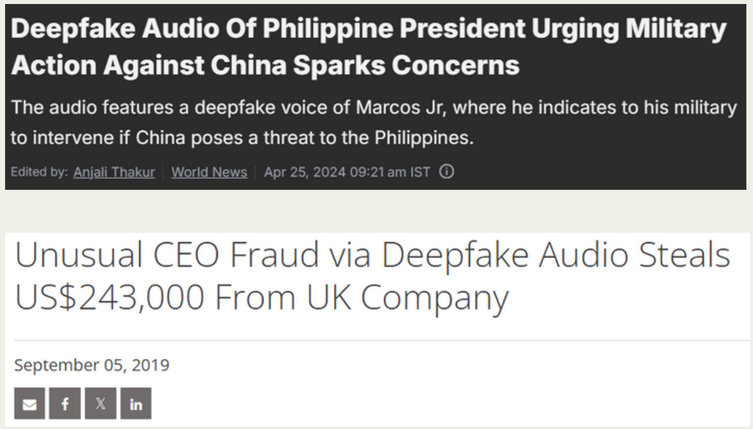
Akanksha Singh
21 August 2025

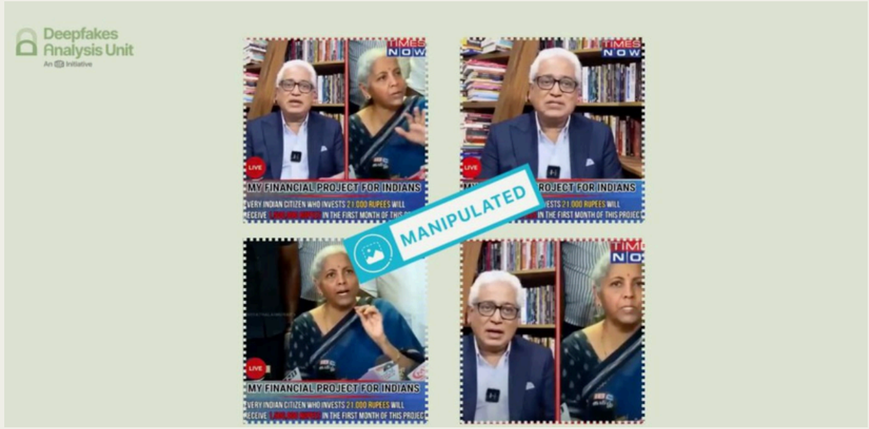
CeRAI
Centre for Responsible AI

MOTIVATION

In discussions with Ms. Pamposh Rain, Head of the Deepfakes Analysis Unit, we found that fact-checkers in India currently grapple with audio spoofs more urgently than visual deepfakes.

- Audio spoof artifacts are **not as conspicuous** as in videos or images.
- Have **perturbations** like background noise, music, and go through compression that makes human and AI detection harder.
- Indian languages are **less represented** compared to the English and Global North.





A near-real audio spoof, threats from audio spoof and DAU in action (Left to Right).

BACKGROUND

Model Name	Feature Type	Input Length	ASVspoof19 eval EER%	ASVspoof19 eval t-DCF	In-the-Wild Data EER%
LCNN	cqtspec	Full	**6.354±0.39**	0.174±0.03	**65.559±11.14**
LCNN	cqtspec	4s	25.534±0.10	0.512±0.00	70.015±4.74
LCNN	logspec	Full	7.537±0.42	**0.141±0.02**	72.515±2.15
LCNN	logspec	4s	22.271±2.36	0.377±0.01	91.110±2.17
LCNN	melspec	Full	15.093±2.73	0.428±0.05	70.311±2.15
LCNN	melspec	4s	30.258±3.38	0.503±0.04	81.942±3.50
LCNN-Attention	cqtspec	Full	**6.762±0.27**	**0.178±0.01**	**66.684±1.08**
LCNN-Attention	cqtspec	4s	23.228±3.98	0.468±0.06	75.317±8.25
LCNN-Attention	logspec	Full	7.888±0.57	0.180±0.05	77.122±4.91
LCNN-Attention	logspec	4s	14.958±2.37	0.354±0.03	80.651±6.14
LCNN-Attention	melspec	Full	13.487±5.59	0.374±0.14	70.986±9.73
LCNN-Attention	melspec	4s	19.534±2.57	0.449±0.02	85.118±1.01
LCNN-LSTM	cqtspec	Full	**6.228±0.50**	**0.113±0.01**	**61.500±1.37**
LCNN-LSTM	cqtspec	4s	20.857±0.14	0.478±0.01	72.251±2.97
LCNN-LSTM	logspec	Full	9.936±1.74	0.158±0.01	79.109±0.84
LCNN-LSTM	logspec	4s	13.018±3.08	0.330±0.05	79.706±15.80
LCNN-LSTM	melspec	Full	9.260±1.33	0.240±0.04	62.304±0.17
LCNN-LSTM	melspec	4s	27.948±4.64	0.483±0.03	82.857±3.49
LSTM	cqtspec	Full	**7.162±0.27**	**0.127±0.00**	**53.711±11.68**
LSTM	cqtspec	4s	14.409±2.19	0.382±0.05	55.880±0.88
LSTM	logspec	Full	10.314±0.81	0.160±0.00	73.111±2.52
LSTM	logspec	4s	23.232±0.32	0.512±0.00	78.071±0.49
LSTM	melspec	Full	16.216±2.92	0.358±0.00	65.957±7.70
LSTM	melspec	4s	37.463±0.46	0.553±0.01	64.297±2.23
MesoNet	cqtspec	Full	**7.422±1.61**	0.219±0.07	54.544±11.50
MesoNet	cqtspec	4s	20.395±2.03	0.426±0.06	65.928±2.57
MesoNet	logspec	Full	8.369±1.06	**0.170±0.05**	**46.939±5.81**
MesoNet	logspec	4s	11.124±0.79	0.263±0.03	80.707±12.03
MesoNet	melspec	Full	11.305±1.80	0.321±0.06	58.405±11.28
MesoNet	melspec	4s	21.761±0.26	0.467±0.00	64.415±15.68
ResNet18	cqtspec	Full	**6.552±0.49**	**0.140±0.01**	**49.759±0.17**
ResNet18	cqtspec	4s	18.378±1.76	0.432±0.07	61.827±7.46
ResNet18	logspec	Full	7.386±0.42	**0.139±0.02**	80.212±0.23
ResNet18	logspec	4s	15.521±1.83	0.387±0.02	88.729±2.88
ResNet18	melspec	Full	21.658±2.56	0.551±0.04	77.614±1.47
ResNet18	melspec	4s	28.178±0.33	0.489±0.01	83.006±7.17
Transformer	cqtspec	Full	**7.498±0.34**	**0.129±0.01**	**43.775±2.85**
Transformer	cqtspec	4s	11.256±0.07	0.329±0.00	48.208±1.49
Transformer	logspec	Full	9.949±1.77	0.210±0.06	64.789±0.88
Transformer	logspec	4s	13.935±1.70	0.320±0.03	44.406±2.17
Transformer	melspec	Full	20.813±6.44	0.394±0.10	73.307±2.81
Transformer	melspec	4s	26.495±1.76	0.495±0.00	68.407±5.53
CRNNSpoof	raw	Full	**15.658±0.35**	**0.312±0.01**	44.500±8.13
CRNNSpoof	raw	4s	19.640±1.62	0.360±0.04	**41.710±4.86**
RawNet2	raw	Full	**3.154±0.87**	**0.078±0.02**	37.819±2.23
RawNet2	raw	4s	4.351±0.29	0.132±0.01	**33.943±2.59**
RawPC	raw	Full	**3.092±0.36**	**0.071±0.00**	**45.715±12.20**
RawPC	raw	4s	3.067±0.91	0.097±0.03	52.884±6.08
RawGAT-ST	raw	Full	**1.229±0.43**	**0.036±0.01**	**37.154±1.95**
RawGAT-ST	raw	4s	2.297±0.98	0.074±0.03	38.767±1.28

Müller et al., “Does Audio Deepfake Detection Generalize?”

Feature type:
Hand-crafted: LFCC/MFCC/CQCC/Mel Spectrogram | Spectrogram
Learning based: Raw waveform, WavLM/Wav2Vec2/Whisper
Metrics: t-DCF, EER, Accuracy (Class/Balanced), F1, AUROC
Data augmentation: SpecAug, RawBoost

MOTIVATION

What Does an Audio Deepfake Detector Focus on? A Study in the Time Domain

Petr Grinberg*†, Ankur Kumar†, Surya Koppisetti†, Gaurav Bharaj†
*EPFL, Switzerland, †Reality Defender Inc., USA
petr.grinberg@epfl.ch, {ankur, surya, gaurav}@realitydefender.ai

relevancy based XAI for transformers ADD models compared against Grad-CAM and SHAP-based methods using faithfulness metrics, perturbation test and partial spoof test,
investigates importance of speech/non-speech, phonetic content and voice onsets/offsets.

A Data-Driven Diffusion-based Approach for Audio Deepfake Explanations

Petr Grinberg[1,2], Ankur Kumar[2], Surya Koppisetti[2], Gaurav Bharaj[2]

[1]EPFL, Switzerland
[2]Reality Defender Inc., USA
petr.grinberg@epfl.ch, {ankur, surya, gaurav}@realitydefender.ai

identifying artifacts by **training a diffusion model** on **spoof spectrogram-difference signal** (vocoded - real audio) as i/p and o/p pairs,
conditioning the model specific to a **classifier** using its **intermediate features** as guidance.

What You Read Isn't What You Hear: Linguistic Sensitivity in Deepfake Speech Detection

Binh Nguyen[1] Shuju Shi[2] Ryan Ofman[3] Thai Le[2]
[1]Independent Researcher [2]Indiana University [3]Deep Media AI
[1]nqbinh17@apcs.fitus.edu.vn [2]{shi16,tle}@iu.edu [3]ryan@deepmedia.ai

adversarial attacks on detectors through **text level manipulations**
signifies model is **sensitive to linguistic complexities**

BASELINES DATASETS

ASVspoof2019_LA
Train
English

ASVspoof2021_DF
Test
English

ASVspoof2021_LA
Test
English

Fake-or-Real
Test
English

In the Wild
Test
English
Collected and not generated

Dataset	Train		Dev		Eval	
	Real	Fake	Real	Fake	Real	Fake
ASVspoof2019_LA	2580 utt	22800 utt	2548 utt	22296 utt	7355 utt	64578 utt
ASVspoof2021_LA	-	-	-	-	~180,000 utt	
ASVspoof2021_DF	-	-	-	-	~600,000 utt	
Fake-or-Real	-	-	-	-	111000 utt	87000 utt
In The Wild	-	-	-	-	20.8 hours	17.2 hours

FURTHER DATASETS

PartialSpoof
Test

ASVspoof but with partial spoofing.

IndieFake
Test

Indian accented English.

MLAAD
Test

Multilingual

HAV-DF
Test

Hindi

Dataset	Train		Dev		Eval	
	Real	Fake	Real	Fake	Real	Fake
PartialSpoof	-	-	-	-	7355 utt	64578 utt
IndieFake	-	-	-	-	8,164	11,396
MLAAD	-	-	-	-		485.3 hours
HAV-DF	-	-	-	-	200	308

MODELS

AASIST (2022)

transformer-based

spectro-temporal attention

Conformer (2024)

transformer+CNN

modelling local (convolutional) and global (transformer)

RawGAT-ST (2021)

graphs+transformer

spectro-temporal graph attention network, fuses spectral and temporal signals within the architecture — rather than combining outputs later, captures cross-domain artifacts

RawNet2 (2021)

raw features

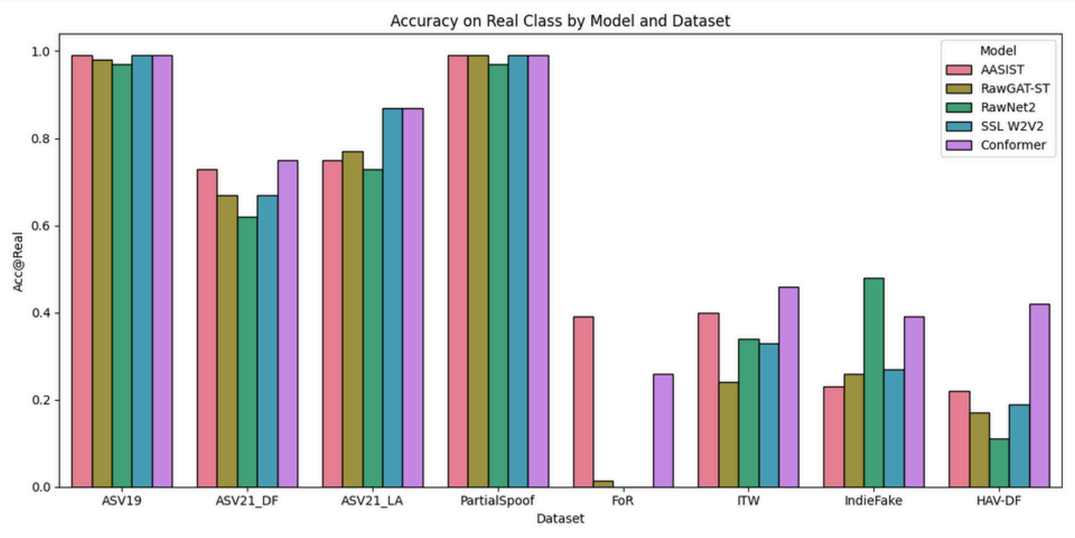
end-to-end raw waveform modelling, captures low-level artifacts that may not appear in handcrafted or spectrogram features

SSL Model wav2vec

self supervised learning

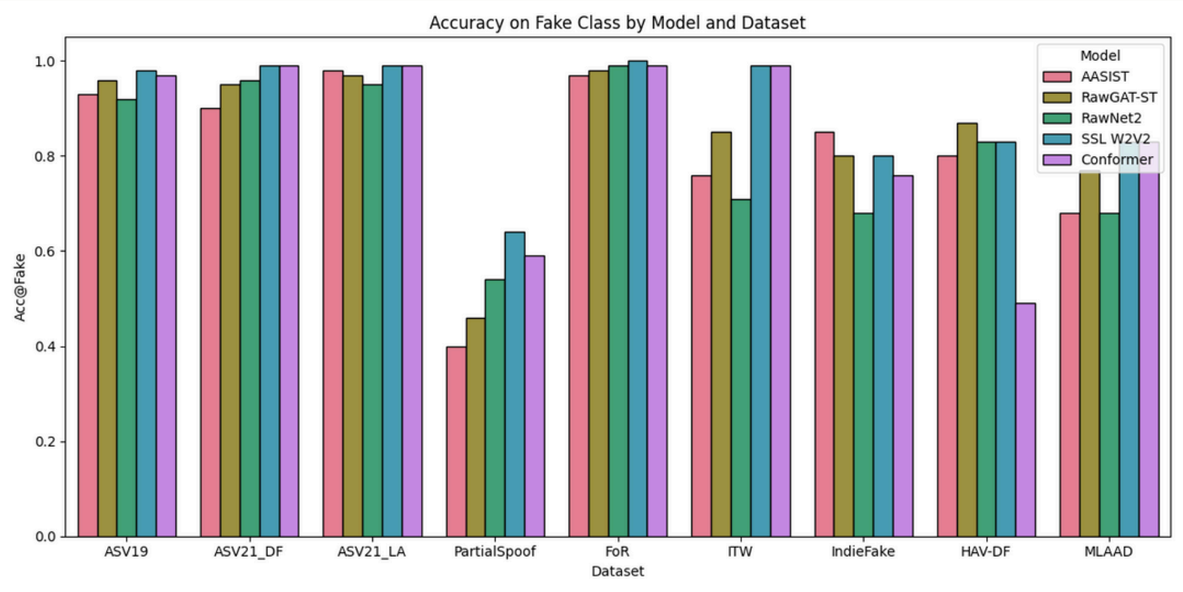
robust and transferable feature extraction

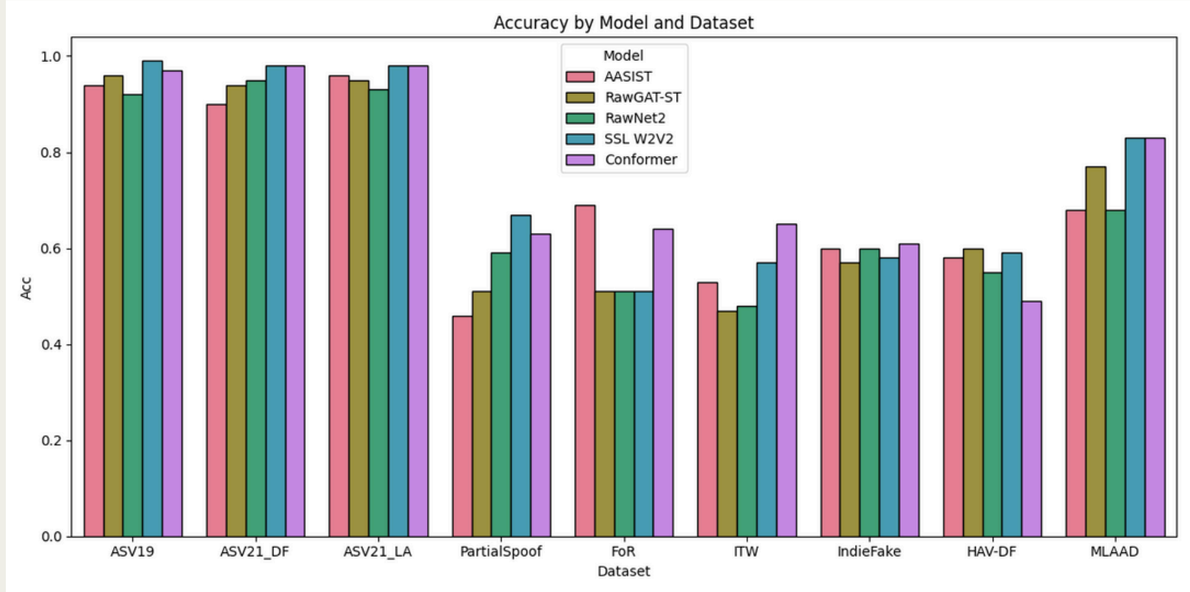
TESTING RESULTS



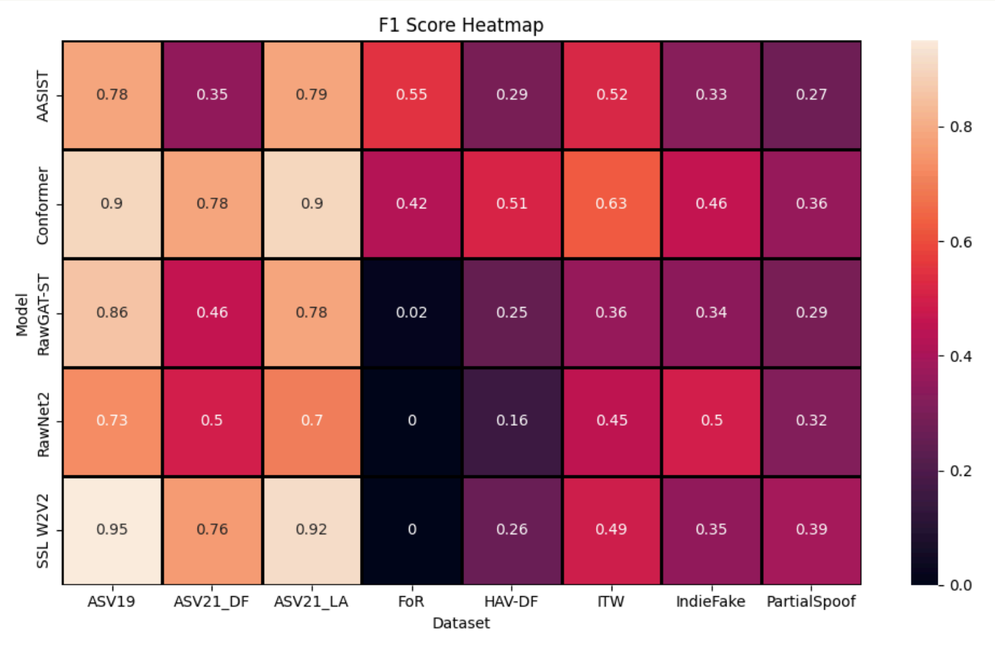
Insights:

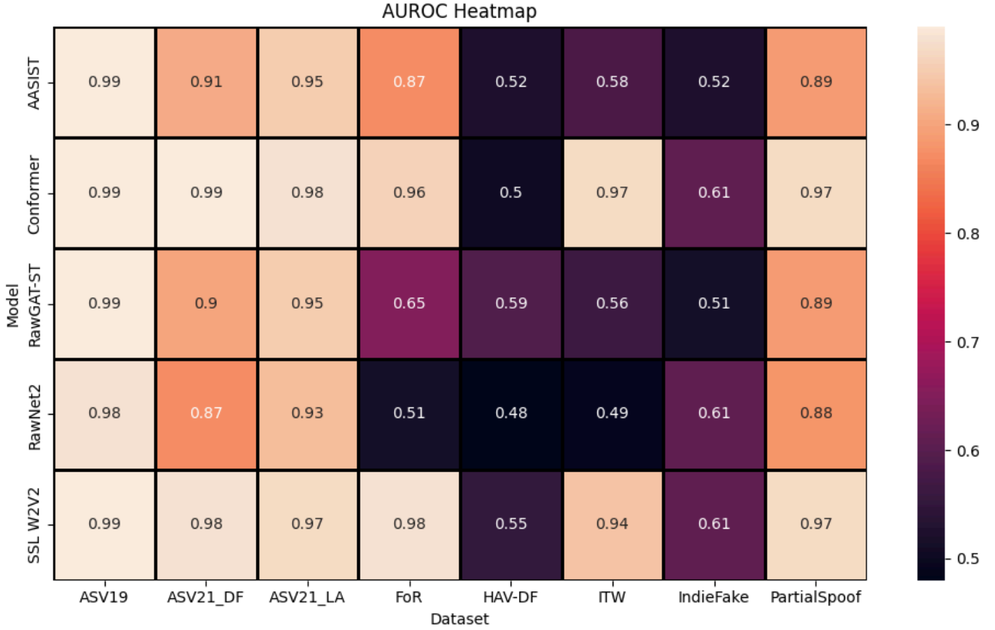
1. **Consistent** change in performance amongst **models.**
2. **Drop** in performance occurs in the case of a **PartialSpoof Fake** class, while the **Real** class remains **unchanged.**
3. **Drop** in performance for **Real** class for all OOD datasets.
4. **OOD language doesn't** necessarily make **detection tougher.**

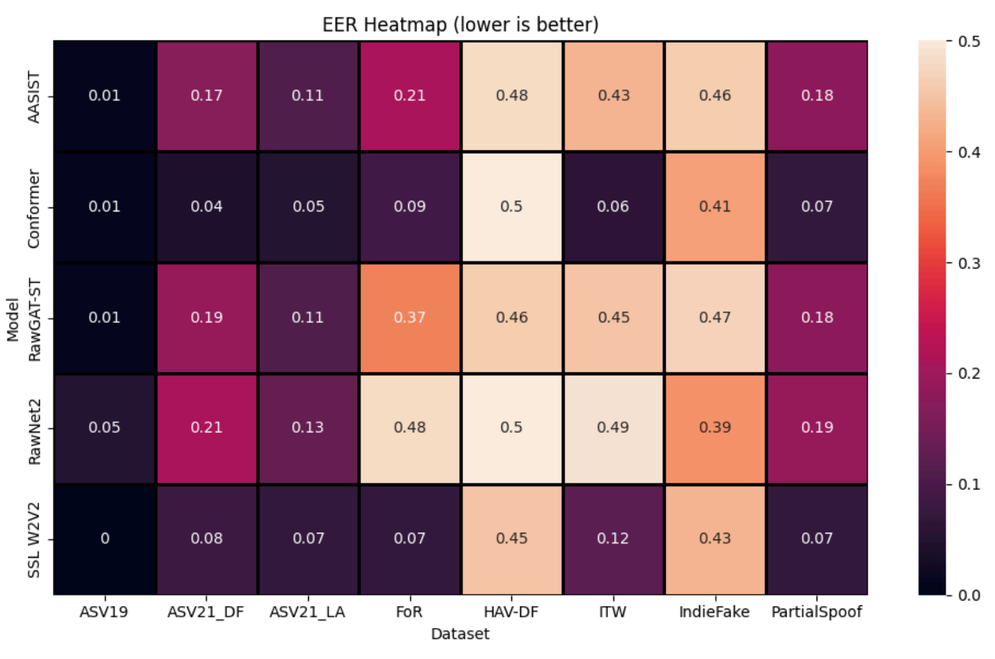




TESTING RESULTS







Insights:

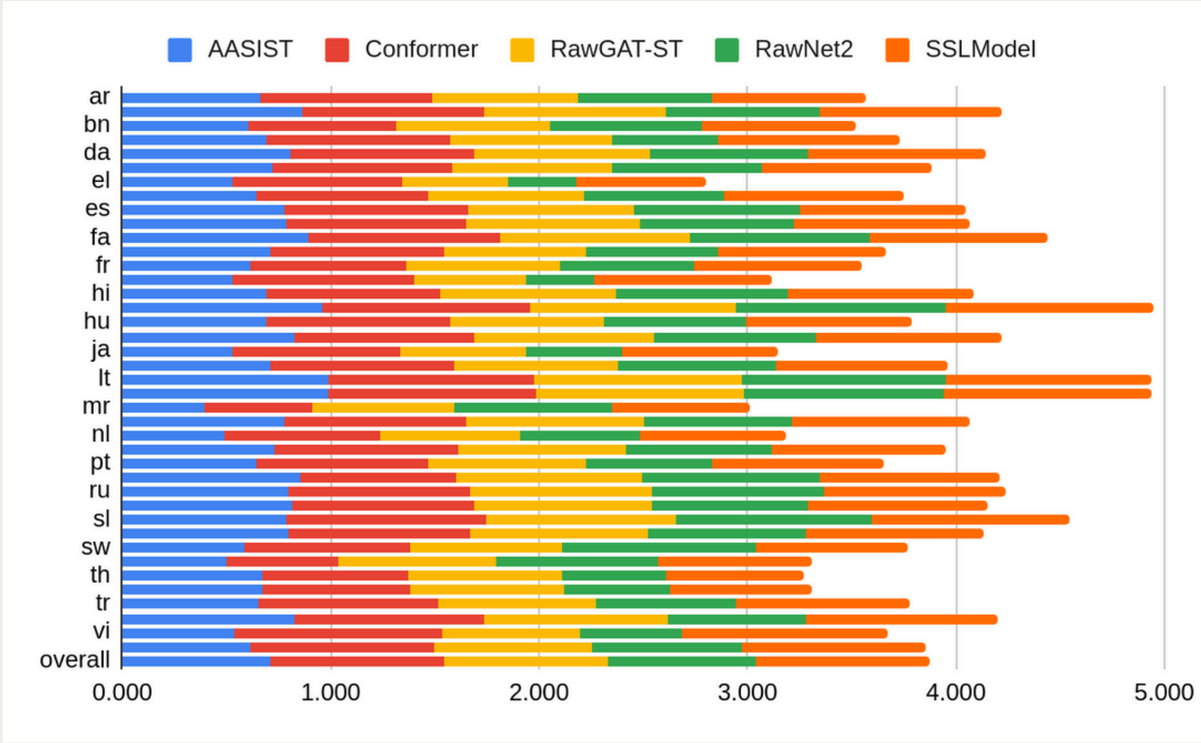
1. Hindi > Indian-accented English > In the wild > Fake or Real

MULTILINGUAL TESTING

Conformer model accuracy for each language

Lang w/ small sample size:
mr: 1000 samples | Acc: 0.517
hr: 1000 samples | Acc: 1.000
lv: 1000 samples | Acc: 1.000

Lang w/ large sample size:
it: 12000 samples| Acc: 0.8691)
fr: 14000 samples | Acc: 0.749
en: 54000 samples | Acc: 0.832



The multilingual benchmark treats each language equally, but accuracy estimates may be less reliable for languages with fewer test samples.

WORK IN PROGRESS & PLAN

Data generation

data to be generated using models

- MMS
- IndicF5
- IndicParler-TTS
- Fastspeech2_HS
- Syspin
- Parrot-TTS
- Coqui-ai
- Indri
- Veena

on following Indic datasets:

- IN22 - AI4Bharat
- Flores+ - Meta

Analysis

- **inference** – Why the change in model performance based on datasets > What is the scope of a model > What case was it developed for – architectural choices and datasets > What are the boundaries of the model?

- **interpretability** – What are the interpretable features for classification?
  - post-hoc interpretability models on top of existing models - gnformation flows by maximising the attention weights, Gradient-based input attribution transformer models, probing – representations from layers and using a simple classifier on top of it.
  -
- **explainability** – How explainable are the results through the features?

Tool creation

ensembling

- based on voting based on merits of models
- using AdaBoost and/or XGBoost.

Thank you!

FLOOR IS OPEN FOR DISCUSSION.

Akanksha Singh
21 August 2025

CeRAI
Centre for Responsible AI