

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



**Cross-Lingual Deepfake Voice Detection Using
Fine-Tuned Self-Supervised Speech
Representations from Raw Waveforms**

by

Imran Javed

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the
Faculty of Computing
Department of Computer Science

2026

Copyright © 2026 by Imran Javed

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.



CERTIFICATE OF APPROVAL

**Cross-Lingual Deepfake Voice Detection Using Fine-Tuned
Self-Supervised Speech Representations from Raw Waveforms**

by

Imran Javed

(MCS241006)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Asif Muhammad	FAST-NU, Islamabad
(b)	Internal Examiner	Dr. M. Masroor Ahmed	CUST, Islamabad

Dr. Aamer Nadeem

Thesis Supervisor

September, 2026

Dr. M. Masroor Ahmed

Head

Dept. of Computer Science

September, 2026

Dr. Muhammad Abdul Qadir

Dean

Faculty of Computing

September, 2026

Author's Declaration

I, **Imran Javed** hereby state that my MS thesis titled "**Cross-Lingual Deepfake Voice Detection Using Fine-Tuned Self-Supervised Speech Representations from Raw Waveforms**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.



(Imran Javed)

Registration No: MCS241006

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Cross-Lingual Deepfake Voice Detection Using Fine-Tuned Self-Supervised Speech Representations from Raw Waveforms**" is solely my research work with no significant contribution from any other person. Small contributions/help wherever taken have been duly acknowledged, and the complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I, as an author of the above-titled thesis, declare that no portion of my thesis has been plagiarized, and any material used as a reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of the MS Degree, the University reserves the right to withdraw/revoke my MS degree, and that HEC and the University have the right to publish my name on the HEC/University website, on which names of students are placed who submitted plagiarized work.



(Imran Javed)

Registration No: MCS241006

Acknowledgement

First and foremost, I would like to thank Allah for granting me the knowledge, strength, and patience needed to complete this research. I am deeply grateful to my supervisor, Dr. Aamer Nadeem, for his guidance, patience, and constructive feedback throughout the course of this thesis.

I would also like to thank the faculty and staff of the Department of Computer Science at Capital University of Science and Technology for their support throughout my degree.

Finally, my sincere thanks go to my family, whose encouragement and support sustained me through every stage of this work.

(Imran Javed)

Abstract

Audio deepfake detection has matured substantially for English speech, but this progress has not extended to most of the world’s languages. Urdu, spoken by well over 230 million people, has almost no dedicated countermeasure research, and prior work shows an English-trained detector applied to Urdu deepfakes performs at chance level. This thesis investigates whether a cross-lingual self-supervised model can be adapted to detect Urdu deepfakes credibly without an Urdu dataset at English-benchmark scale, and whether the transfer failure is symmetric in the reverse direction.

The proposed system fine-tunes XLS-R 300M directly on raw audio waveforms, freezing the lower fourteen of twenty-four transformer layers and fine-tuning the upper ten together with a learnable weighted-sum aggregation and a lightweight classification head. Training uses class-weighted loss, voice-activity-based silence trimming, and sliding-window segmentation that expands the Urdu training pool by 85 percent. Two datasets, ASVspoof 2019 LA (English) and the CSALT Urdu Deepfake Audio Dataset, are combined into a six-experiment protocol covering intra-dataset baselines, both directions of zero-shot cross-lingual transfer, and bi-lingual training.

Results show near-ceiling intra-dataset detection, with an Equal Error Rate of 0.58% for English and 3.89% for Urdu. Zero-shot cross-lingual transfer degrades severely in both directions, reaching 24.56% (English-trained tested on Urdu) and 27.07% (Urdu-trained tested on English). Combining English and Urdu training data recovers most of this loss, reducing the cross-lingual Equal Error Rate on Urdu by 57.3% relative to the zero-shot condition, at a cost of only 1.37 percentage points to English-only accuracy. A modest quantity of target-language data, combined with a principled partial fine-tuning strategy, can therefore substantially narrow the cross-lingual detection gap for a low-resource language pair without large-scale data collection.

Contents

Author’s Declaration	iii
Plagiarism Undertaking	iv
Acknowledgement	v
Abstract	vi
List of Figures	x
List of Tables	xi
Abbreviations	xii
1 Introduction	1
1.1 Background	1
1.2 The Impact of Deepfakes	3
1.3 Relevant Prior Work in Brief	6
1.4 Research Motivation	7
1.5 Problem Statement	8
1.6 Research Objectives	8
1.7 Research Questions	9
1.8 Scope and Delimitations	9
1.9 Thesis Organisation	10
2 Literature Review	11
2.1 Relevant Prior Work	12
2.2 Foundational Benchmarks: The ASVspoof Lineage	14
2.3 The XLS-R Backbone and Layer-Wise Representation Structure	15
2.3.1 XLS-R and wav2vec 2.0	16
2.4 Other Self-Supervised Front-Ends for Spoof Detection	16
2.5 Cross-Lingual and Multilingual Generalisation	17
2.6 Low-Resource Language Datasets and Detection: Urdu and Beyond	19
2.7 Class Imbalance and Training Dynamics	20
2.8 Domain Shift, Codec Robustness and Input Representation	21

2.8.1	Raw Waveform Versus Spectrogram Input	22
2.9	Surveys and Expert Assessments	22
2.10	The Generative Side: Why Cross-Lingual Detection Keeps Getting Harder	23
2.11	Most Recent Advances (2025–2026)	24
2.12	Methodology Comparison	26
2.13	Datasets Used: Comparative Table	29
2.14	Summary of Research Gaps	31
3	Proposed Methodology	32
3.1	Model Architecture: XLS-R 300M	34
3.1.1	How XLS-R Works: A Multilingual Speech Recognition Model	34
3.1.2	Cross-Lingual Transfer	34
3.1.3	How the Model Extracts Features and Represents Language	35
3.1.3.1	Raw Audio to Local Features (CNN Encoder)	35
3.1.3.2	Turning Sound into Latent Speech Units (Quantizer)	36
3.1.3.3	Learning Context with the Transformer and Contrastive Loss	36
3.1.3.4	How Does the Model Identify a Language	36
3.2	Dataset Collection and Characterisation	37
3.2.1	ASVspoof 2019 Logical Access Partition	37
3.2.2	CSALT Urdu Deepfake Detection Dataset	38
3.3	Audio Preprocessing and Segmentation	40
3.3.1	Channel Normalisation and Resampling	41
3.3.2	Voice Activity Detection	41
3.3.3	Amplitude Normalisation	41
3.3.4	Sliding-Window Segmentation	42
3.3.5	Why the Uppermost Ten Layers Were Selected for Fine-Tuning	43
3.3.6	Learnable Weighted Layer Aggregation	44
3.3.7	Temporal Pooling and Classification Head	45
3.4	Class Imbalance and Role of Class Weighting	45
3.4.1	Why Class Weights Were Necessary	46
3.4.2	How the Weights Were Computed	47
3.4.3	Observed Impact of Class Weighting	48
3.5	Evaluation Metrics	48
3.5.1	Equal Error Rate	48
3.5.2	Minimum Tandem Detection Cost Function	49
3.5.3	Area Under the ROC Curve	50
3.5.4	Accuracy, Precision, Recall, F1-Score and Confusion Matrices	50
4	Results and Discussion	51
4.1	Training Configuration	52
4.2	Evaluation Protocol: A Six Experiment Cross Lingual Matrix	53
4.2.1	Data Splitting and Leakage Prevention	54
4.2.2	Rationale for Reusing Models Across Experiments	54

4.3	Intra-Dataset Baselines	55
4.3.1	English Intra-Dataset (EN→EN)	56
4.3.2	Urdu Intra-Dataset (UR→UR)	57
4.3.3	Interpreting the Baseline Gap	57
4.4	Zero-Shot Cross-Lingual Transfer	58
4.4.1	English Trained, Urdu Tested (EN→UR)	58
4.4.2	Urdu Trained, English Tested (UR→EN)	59
4.4.3	The Cross-Lingual Asymmetry	60
4.5	Bi-Lingual Training	61
4.5.1	Bi-Lingual Training, Urdu Test Set	61
4.5.2	Bi-Lingual Training, English Test Set	62
4.5.3	The Favourable Trade-off	63
4.6	Comparative Analysis Across All Six Experiments	64
4.6.1	Equal Error Rate	64
4.6.2	Area Under the ROC Curve	65
4.6.3	Accuracy and F1-Score	65
4.6.4	Precision and Recall	67
4.7	The Cross-Lingual Generalisation Pattern	67
4.8	Qualitative Spectrogram Analysis	68
4.9	Comparison with Published Literature	70
4.9.1	Comparison with the Urdu Dataset Baseline	70
4.9.2	Comparison with DECRO	71
4.9.3	Comparison with the Broader Cross-Lingual Literature	71
4.9.4	Comparison with XLS-R plus SLS	72
4.10	Summary of Findings	72
5	Conclusion and Future Work	74
5.1	Contributions	74
5.2	Limitations	75
5.2.1	Synthesis-Method Coverage	75
5.2.2	Corpus Scale Asymmetry	75
5.2.3	Segment-Level Rather Than Utterance-Level Evaluation	76
5.2.4	Two-Language Scope	76
5.3	Future Work	77
5.3.1	Broader Synthesis-Method Coverage	77
5.3.2	Real-Time and Deployment-Oriented Evaluation	77
5.3.3	Adversarial Robustness	77
5.3.4	Explainability and Fairness Analysis	78
5.4	Conclusion	78
	Bibliography	79

List of Figures

2.1	Timeline of principal benchmarks and datasets.	12
2.2	Reported EER degradation from within-language or within-corpus to cross-lingual or cross-domain evaluation conditions.	14
3.1	End-to-end architecture of the proposed cross-lingual deepfake voice detection system, from raw waveform input through the XLS-R encoder to the final real or fake decision.	33
3.2	Feature extraction and multilingual representation pipeline in XLS-R.	35
3.3	The audio preprocessing and segmentation pipeline.	40
4.1	(EN→EN): English intra-dataset classification metrics.	56
4.2	(UR→UR): Urdu intra-dataset classification metrics.	57
4.3	(EN→UR): cross-lingual Urdu tested classification metrics.	59
4.4	(UR→EN): cross-lingual English tested classification metrics	60
4.5	(EN+UR→UR): bi-lingual training, Urdu test classification metrics.	62
4.6	(EN+UR→EN): bi-lingual training, English test classification metrics.	63
4.7	Equal Error Rate across all six experiments (lower is better).	64
4.8	Area Under the ROC Curve across all six experiments, with the random-classifier reference line at $AUC = 0.5$ (higher is better).	65
4.9	Classification accuracy across all six experiments, evaluated at each experiment's EER threshold.	66
4.10	F1-score across all six experiments, evaluated at each experiment's EER threshold.	66
4.11	Precision and recall, evaluated at each experiment.	67
4.12	Cross-lingual generalisation summary.	68
4.13	Mel-spectrogram comparison of representative bonafide and spoofed segments for both languages.	69

List of Tables

2.1	Comparison of model architectures, datasets, and reported results across principal studies reviewed.	26
2.2	Datasets referenced across the reviewed literature, along with languages details and publicly availability.	29
3.1	Summary statistics for both corpora before and after preprocessing, together with the combined pool used for bi-lingual training experiments.	39
3.2	Class weights applied to the CrossEntropyLoss function.	47
4.1	The six-experiment evaluation protocol. Experiments 5 and 6 evaluate existing models from Experiments 4 and 3 respectively on a second held-out test set, rather than requiring separate training runs, since the underlying training data and model weights are identical in each pair.	53
4.2	Complete metric summary across all six experiments. The lowest precision value observed anywhere in the study occurs in Experiment 5, flagging the most severe individual metric collapse.	55
4.3	Raw confusion-matrix counts underlying Table 4.2, evaluated at each experiment’s own Equal-Error-Rate threshold. “Fake” and “Real” refer to the bonafide/spoof ground truth and predicted label respectively.	55

Abbreviations

AASIST	Audio Anti-Spoofing using Integrated Spectro-Temporal graph attention networks
AMP	Automatic Mixed Precision
ASV	Automatic Speaker Verification
AUC	Area Under the (ROC) Curve
CSALT	Center for Speech and Language Technologies (LUMS)
DECRO	Cross-lingual deepfake evaluation dataset (English-Mandarin)
EER	Equal Error Rate
FAR	False Acceptance Rate
FRR	False Rejection Rate
F1-Score	F1-Score (harmonic mean of precision and recall)
HuBERT	Hidden-Unit BERT (self-supervised speech model)
min-tDCF	Minimum tandem Detection Cost Function
MLAAD	Multi-Language Audio Anti-Spoofing Dataset
ROC	Receiver Operating Characteristic
SLS	Sensitive Layer Selection
SSL	Self-Supervised Learning
TTS	Text-to-Speech
VAD	Voice Activity Detection
VC	Voice Conversion
VITS	Variational Inference with adversarial learning for end-to-end Text-to-Speech
WavLM	Wave Language Model (self-supervised speech model)
XLS-R	Cross-Lingual Speech Representations (multilingual wav2vec 2.0 variant)

Chapter 1

Introduction

1.1 Background

A deepfake is synthetic media, image, video, or audio, generated or manipulated by a machine learning model so that it depicts a person saying or doing something they never actually said or did, in a way convincing enough to pass for genuine. The term originated around 2017 from a Reddit contributor who used deep neural networks to superimpose faces onto video, and it has since broadened to cover any AI-generated forgery of a real person’s likeness or voice. Audio deepfakes, the specific concern of this thesis, are synthetic speech waveforms produced by text-to-speech (TTS) or voice-conversion (VC) systems trained to imitate a target speaker’s voice, timbre, prosody, and speaking style closely enough that a listener, or an automated system, mistakes the output for a genuine recording of that person.

Understanding what an audio signal actually is, and which of its properties a detector can exploit, is useful before discussing how such deepfakes are produced and detected. An audio signal is, physically, a continuous variation in air pressure over time produced by a vibrating source such as the human vocal folds; a microphone converts this pressure wave into a continuous (analog) electrical signal, which is then digitised for computational processing by sampling it at discrete

time intervals and quantising each sample’s amplitude to a finite set of levels, typically 16,000 samples per second at 16-bit resolution for the speech corpora used in anti-spoofing research [1]. The resulting digital waveform is characterised by several properties relevant to detection: its amplitude (loudness, the height of the waveform at a given instant), its frequency content (pitch and timbre, how rapidly amplitude oscillates, conventionally analysed across the roughly 20 Hz to 20 kHz range of human hearing), its phase, and its duration. For speech specifically, this raw pressure waveform simultaneously carries phonemic content (the linguistic message), prosodic structure (pitch contour, rhythm, and stress patterns), and speaker-specific timbral characteristics, all of which a text-to-speech or voice-conversion system must reproduce convincingly and all of which a countermeasure can, in principle, draw on to distinguish genuine from synthetic speech [2]. Detection systems can operate on this waveform in one of two broad ways: by first transforming it into a frequency-domain representation such as a spectrogram or mel-frequency cepstral coefficients, which discards some raw temporal detail in the process, or by operating end-to-end directly on the raw time-domain waveform, an approach shown to preserve fine-grained synthesis artefacts that handcrafted spectral features can smooth over [3–5]. This thesis adopts the latter, raw-waveform approach, for reasons detailed in Chapter 2 and Chapter 3.

The technology underlying voice cloning has progressed from requiring hours of clean, studio-quality recordings of a target speaker as recently as 2019, to requiring only a few seconds of reference audio by 2023, a shift driven by large neural TTS architectures such as Tacotron 2, VITS, and, more recently, zero-shot diffusion- and flow-matching-based cloning systems trained on massive multilingual speech corpora [6, 7]. This trajectory matters because it has moved voice cloning from a specialist capability, requiring machine learning expertise, GPU compute, and substantial target-speaker audio, to something achievable by a non-expert with a smartphone app and a short public recording, a shift documented across the generative-side literature reviewed in Chapter 2. The consequence is a rapidly expanding attack surface. Any person whose voice has been recorded and made public, on a podcast, in a news interview, on social media, in a company earnings

call, is now a plausible target for synthetic impersonation.

Detection research has struggled to keep pace with this generative progress for a structural reason as much as a technical one. Countermeasure systems are trained and evaluated almost exclusively on English-language benchmarks, principally the ASVspoof challenge series [1, 8], while the generative systems attackers can deploy are increasingly multilingual by default, trained on hundreds of languages simultaneously as a side effect of how large speech foundation models such as XLS-R are built [9]. This asymmetry, well-resourced, English-centric detection research facing an increasingly language-agnostic threat, is the starting point for this thesis, and Section 1.3 returns to it as the specific problem this work addresses.

1.2 The Impact of Deepfakes

Voice deepfake detection is motivated by harms that are already documented, not hypothetical, and that span financial, political, personal, legal, and institutional domains simultaneously.

The clearest and earliest-documented harm is direct financial fraud. In 2019, the Wall Street Journal reported that fraudsters used AI-generated speech to impersonate a German energy company’s CEO, instructing a UK subsidiary to wire EUR 220,000 (roughly USD 243,000) to a Hungarian supplier account [10], a transfer the victim authorised because he recognised his superior’s accent and cadence over the phone. Similar fraud has since been reported against a Hong Kong bank, where a cloned voice matching a company director deceived a branch manager into authorising USD 35 million in transfers [11], and against private individuals through “grandparent scams” that clone a family member’s voice from a few seconds of social-media audio to request emergency funds. Consumer-protection reporting since 2023 describes voice-cloning scams as among the fastest-growing categories of imposter fraud, with aggregate losses estimated in the hundreds of millions annually. Financial institutions face a related second-order cost, since voice biometric authentication, marketed for years as a security convenience, is

directly undermined by the same cloning technology, forcing costly re-engineering of authentication pipelines.

Deepfake audio also threatens democratic processes, since recorded speech needs no visual likeness to be convincing and can spread cheaply through robocalls and messaging platforms. In January 2024, a robocall using an AI-cloned voice of then-US President Joe Biden reached thousands of New Hampshire voters urging them not to vote, prompting an FCC ruling that AI-generated robocall voices fall under existing anti-robocall law [12, 13]. Fabricated political audio has also circulated during Slovakia’s 2023 election [14] and across South Asian election cycles, a trend of direct relevance given this thesis’s South Asian language focus. Because the gap between a fabricated clip’s release and its debunking is often measured in hours, while voters may form impressions within minutes, detection capability functions as part of electoral integrity infrastructure rather than a purely academic concern.

Away from these headline cases, the most pervasive harm is interpersonal: harassment and psychological abuse directed at private individuals with no institutional resources to respond. Documented patterns include “virtual kidnapping” extortion using a cloned family member’s voice, intimate-partner harassment using fabricated recordings, and cloned voices used against children from a few seconds of public video. Survivor accounts describe a lasting erosion of trust in phone calls from known contacts that persists even among people never personally targeted. A related, gendered harm concerns non-consensual synthetic intimate content, which has prompted legislative responses including provisions in the UK’s Online Safety Act and several US state laws enacted since 2023.

Audio recordings have long served as evidence in legal proceedings, and convincing synthetic speech threatens this role in two directions at once: fabricated audio can be introduced as false evidence, and, conversely, the so-called “liar’s dividend” lets genuine incriminating recordings be dismissed as fabrications once fabrication is known to be possible. Both pressures push forensic audio authentication toward needing to scale, motivating the kind of automated countermeasures this thesis contributes to, even though courtroom-grade authentication demands a far higher

standard than the research-benchmark accuracy reported in Chapter 4. Beyond the courtroom, organisations face business email compromise attacks that add a synthetic-voice element to a fraudulent request, and formal voice biometric systems used by banks and telecom providers face direct exposure, since a successfully cloned voice can, in principle, defeat a voice-based access check outright.

News organisations rely on recorded audio as primary evidentiary material, and convincing deepfakes create risk in both directions here too: newsrooms may unknowingly broadcast fabricated audio, or genuine recordings may be dismissed as fabricated by the people they implicate, a tactic press-freedom organisations have flagged as a tool for discrediting investigative journalism in repressive contexts. At the state level, defence and intelligence communities have flagged synthetic audio as a disinformation tool, including fabricated recordings of military commanders during active conflicts. Because the marginal cost of generating synthetic audio in a new language has fallen sharply with massively multilingual text-to-speech systems, this threat is no longer confined to a handful of high-resource languages, a point of direct relevance to the low-resource-language focus of this thesis: a detection capability that exists only for English or Mandarin offers no protection to the majority of the world’s population, including speakers of Urdu, spoken natively or as a second language by well over 230 million people across Pakistan and the wider South Asian diaspora.

A final category of harm concerns voice actors and public figures whose vocal likeness carries commercial and creative value independent of any fraudulent use. Unauthorised voice cloning has generated disputes in the entertainment industry and prompted union-level labour action, including provisions addressing AI voice replication in the 2023 SAG-AFTRA strike agreement [15]. This category is commercial and consent-based rather than fraudulent, but it illustrates that the technology’s impact reaches even domains where no listener is being deceived about authenticity, only the person whose voice was cloned was never asked.

Taken together, these harms make clear that voice deepfake detection is not a narrow technical curiosity but a cross-cutting requirement touching financial systems,

electoral integrity, personal safety, judicial processes, corporate security, press freedom, national security, and creative labour rights simultaneously. The remainder of this chapter narrows this broad motivation down to the specific, tractable research problem this thesis addresses, namely the near-total absence of deepfake detection capability for a language spoken by hundreds of millions of people, and how that gap might be narrowed without requiring a dataset at the scale of the field’s English-language benchmarks.

1.3 Relevant Prior Work in Brief

Chapter 2 reviews the literature in full, but the shape of the gap this thesis addresses is worth summarising here. Cross-lingual audio deepfake detection has been studied directly by only a handful of papers. Ba et al. [16] establish that an English-trained detector degrades substantially on Mandarin and show that adversarial domain adaptation, which explicitly trains a model to discard language-identifying information, recovers much of this loss for that language pair.

Munir et al. [17] document the same failure for Urdu specifically, finding that an English-trained detector performs at chance level on Urdu deepfakes, but stop short of testing any adaptation strategy at all. Separately, a body of work on multilingual self-supervised encoders, most relevantly XLS-R [9], shows that pre-training breadth across languages improves cross-lingual robustness generally [18], without testing this claim for any specific low-resource language pair.

Two challenges recur across this literature. The first is data scarcity: low-resource languages such as Urdu have nothing resembling the roughly 121,000-utterance scale of the English ASVspoof benchmark [1], and collecting a comparable corpus from scratch is impractical for most research groups. The second is that multilingual pretraining alone does not guarantee cross-lingual transfer, as the Mandarin and Urdu failures above both demonstrate. Proposed solutions in the literature include target-language fine-tuning [19], and combining data from multiple languages during training [20], but despite the scale no published study has tested

any of these solutions for the English–Urdu pair specifically, or compared both directions of transfer between them.

This thesis is positioned directly against that gap. It asks whether a comparatively simple solution, joint fine-tuning on combined English and Urdu data, can recover a meaningful share of the cross-lingual performance loss for this language pair, without the adversarial machinery Ba et al. [16] required for Mandarin, and without collecting an Urdu corpus at English-benchmark scale.

1.4 Research Motivation

Three considerations motivate this specific investigation beyond the general case for deepfake detection research made in Section 1.2.

First, Urdu’s absence from the countermeasure literature is not a marginal gap. It is a language with more native and second-language speakers than the United Kingdom, France, Germany, and Italy combined, spoken across a region already documented as a target for political audio disinformation, and served by essentially no automated protection against synthetic voice fraud at the time of writing.

Second, the self-supervised speech modelling literature reviewed in Chapter 2 suggests, but does not directly demonstrate, that a cross-lingual model such as XLS-R, pretrained on Urdu among 128 languages [9], should be able to transfer detection capability to Urdu more efficiently than training a detector from scratch, an efficiency this thesis sets out to quantify rather than assume.

Third, the specific combination of design choices this thesis tests, partial layer fine-tuning grounded in interpretability research, class-weighted training to handle the field’s characteristic real/fake imbalance, and a raw-waveform rather than spectrogram-based pipeline, has not previously been assembled and evaluated together for the English–Urdu language pair, making the resulting empirical findings a genuine contribution rather than a straightforward replication of an existing recipe on new data.

1.5 Problem Statement

Automated deepfake voice detection has matured substantially for English speech, driven by a well-resourced benchmark ecosystem, principally the ASVspoof challenge series [1, 8], and, more recently, large self-supervised speech models such as XLS-R that reduce English detection error rates to a fraction of one percent when properly fine-tuned [21]. This progress has not extended to the Urdu. A language with a phoneme inventory, script, and prosodic structure substantially different from English, and spoken by well over 230 million people, has limited dedicated countermeasure research. At the time this thesis was undertaken, exactly one peer-reviewed, purpose-built Urdu deepfake dataset existed in the literature [17], and the same study that introduced it found that an English-trained detector applied to Urdu deepfakes performed at chance level, indistinguishable from a coin flip, while human listeners fared only marginally better [17].

1.6 Research Objectives

Prior work, reviewed in Section 1.3, already shows that strong intra-dataset detection is achievable for English, using XLS-R and other architectures respectively. Establishing a matching baseline under this thesis’s own architecture is therefore treated as a necessary methodological step, reported in Chapter 4, rather than a standalone research objective. Against that background, this research pursues three objectives, each addressed through a distinct component of the experimental design described in Chapter 3.

- i. To evaluate cross-lingual transfer directly and in both directions, rather than relying on the untested assumption that multilingual pretraining alone guarantees cross-lingual transfer, by measuring detection performance for English-trained models tested on Urdu and Urdu-trained models tested on English, under a controlled, leakage-free evaluation protocol.

- ii. To identify and apply an architecture and training configuration, specifically partial-layer fine-tuning, class-weighted loss, and end-to-end raw-waveform input, that reduces computational cost relative to full-model fine-tuning while preserving or improving detection performance, so that the resulting approach remains practical for research groups without large-scale compute budgets.
- iii. To quantify whether, and by how much, combining a modest quantity of target-language (Urdu) training data with a larger majority-language (English) pool narrows the cross-lingual performance gap identified under Objective 1, and at what cost, if any, to majority-language detection performance.

1.7 Research Questions

- i. Does detection performance change under zero-shot cross-lingual transfer between English and Urdu?
- ii. Does the proposed partial fine-tuning and class-weighted training strategy improve competitive detection performance?
- iii. Does combining English and Urdu training data reduce the cross-lingual error?

1.8 Scope and Delimitations

This research is deliberately bounded in several respects, each stated here to set correct expectations for the results in Chapter 4 and the limitations discussed in Chapter 5. The study covers exactly two languages, English and Urdu, and does not claim its findings generalise automatically to other language pairs, particularly pairs not well represented in XLS-R’s pretraining corpus. The Urdu spoofed data available for this study was generated by two text-to-speech systems, Tacotron 2 and VITS. Newer zero-shot voice-cloning systems, discussed as an emerging threat

in Section 1.2, are not represented in either training or test data. The English data is drawn from a single benchmark corpus, ASVspoof 2019 LA. Broader English resources such as In-the-Wild or MLAAD, discussed in Chapter 2, were not incorporated into training. Finally, this thesis addresses detection performance under controlled, held-out evaluation conditions.

1.9 Thesis Organisation

The remainder of this thesis is organised as follows. Chapter 2 reviews the published literature on audio deepfake detection, self-supervised speech representation learning, and cross-lingual and low-resource-language countermeasure research, establishing the specific gap this thesis addresses. Chapter 3 details the proposed methodology. Dataset collection and preprocessing, the XLS-R-based detection architecture and the reasoning behind its specific fine-tuning strategy, the training configuration including the treatment of class imbalance, and the six-experiment cross-lingual evaluation protocol. Chapter 4 presents and discusses the results of all six experiments, including per-experiment confusion matrices and classification metrics, comparative analysis across experiments, a qualitative spectrogram-based analysis, and an honest account of the threats to validity affecting the findings. Chapter 5 concludes the thesis, summarising the contributions made, and sets out directions for future work that extend beyond this thesis’s stated scope. A consolidated reference list follows Chapter 5.

Chapter 2

Literature Review

This chapter reviews the published research on audio deepfake detection most relevant to a cross-lingual, self-supervised-learning-based approach targeting English and Urdu speech. The review pool was assembled from IEEE, ACM, ACL, EMNLP, and Interspeech proceedings, Springer and Elsevier journals, and arXiv preprints later accepted at one of these venues, spanning 2020 through early 2026, with a deliberate bias toward the last two years given how quickly the field is moving. Forty nine sources are cited in total, drawn from foundational anti-spoofing benchmarks, self-supervised representation learning, cross-lingual and multilingual generalisation studies, low-resource-language datasets, domain and codec robustness research, class-imbalance and training-dynamics studies, and recent survey articles.

The chapter is organised thematically. Section 2.1 introduces the three studies considered most directly relevant to the approach evaluated later in this thesis. Section 2.2 covers the foundational English-only benchmarks that underpin nearly all subsequent work. Section 2.3 examines the XLS-R self-supervised backbone and the layer-wise interpretability literature that explains what different parts of such a model encode. Section 2.4 surveys other self-supervised front-ends proposed for spoof detection. Section 2.5 covers cross-lingual and multilingual generalisation research specifically. Section 2.6 focuses on low-resource-language datasets,

including Urdu. Section 2.7 examines class imbalance and training-dynamics research. Section 2.8 covers domain, codec, and real-world generalisation, including the choice between raw-waveform and spectrogram-based input representations. Section 2.9 reviews field-level surveys and expert assessments. Section 2.10 discusses the generative side of the problem, why detection keeps getting harder as synthesis technology improves. Section 2.11 isolates the most recent (2025–2026) contributions. Sections 2.12 through 2.14 present comparative tables of methodology, datasets, and results across the studies reviewed, and Section 2.14 closes the chapter by summarising the research gaps that emerge across it.

Figure 2.1 plots the principal benchmarks and datasets discussed throughout this chapter along a single timeline spanning 2019 to 2026, giving a chronological anchor for the thematic discussion that follows.

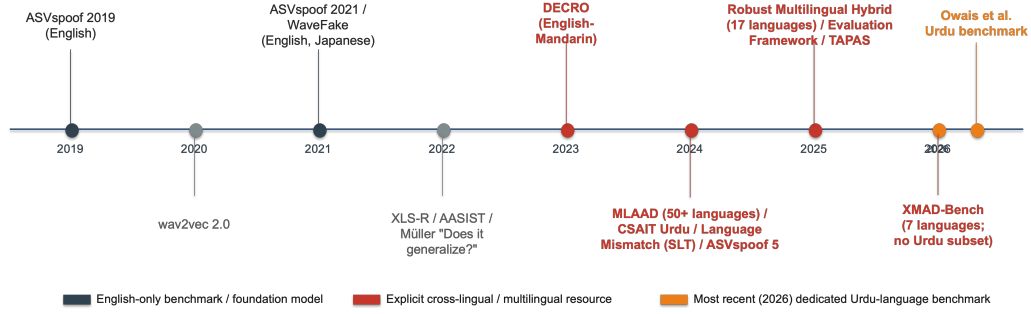


FIGURE 2.1: Timeline of principal benchmarks and datasets.

2.1 Relevant Prior Work

Three studies stand out from the wider literature reviewed in this chapter as the most directly relevant to the approach evaluated here, each supplying a different piece of the problem this thesis addresses.

Munir et al. [17], “Constructing and Evaluating a Specialized Urdu Deepfake Audio Dataset”, is the earliest peer-reviewed, purpose-built English to Urdu deepfake resource identified in this review. It introduces the CSALT Urdu Deepfake Audio Dataset, built from Tacotron and VITS synthesised Urdu speech paired with bona fide recordings, and evaluates a pretrained AASIST-L countermeasure

trained only on English ASVspoof data against it. The authors report an Equal Error Rate of 49.5 percent on VITS spoofed audio and 52.4 percent on Tacotron spoofed audio, both close to chance level, and a companion human-listening study found ordinary listeners distinguished genuine from synthetic Urdu speech only slightly better than chance (ROC AUC of 0.63). The paper documents this transfer failure clearly but does not attempt to close it: it does not fine-tune a model on Urdu, does not test whether combining Urdu data with a larger English pool improves detection, and does not use a raw-waveform self-supervised encoder.

Ba et al. [16], “Transferring Audio Deepfake Detection Capability across Languages”, constructs DECRO, a paired English to Mandarin cross-lingual evaluation dataset of comparable duration in each language. Its central contribution is a bidirectional evaluation design, measuring detection accuracy for an English-trained model tested on Mandarin and a Mandarin-trained model tested on English, rather than only one direction, and comparing naive joint training, straightforward target-language fine-tuning, and an adversarial domain adaptation method that explicitly discards language-identifying information while retaining real-versus-fake signal. The adversarial approach proves strongest of the three. This bidirectional design is the closest existing template for the six-experiment protocol adopted in Chapter 3, although the adversarial mechanism itself is not the adaptation strategy tested here.

Zhang et al. [21], “Audio Deepfake Detection with Self-Supervised XLS-R and SLS Classifier”, pairs the XLS-R self-supervised encoder with a sensitive layer selection classifier that learns a weighting over the model’s 24 transformer layers rather than reading out only the final layer. This is the strongest single-model English result identified anywhere in this review, reducing Equal Error Rate on ASVspoof 2019 from a prior state of the art of 4.07 percent to 1.15 percent, and the learnable weighted-sum aggregation described in Chapter 3 directly adapts this layer-selection idea. Zhang et al. evaluate exclusively on English and do not test cross-lingual transfer, and fine-tune XLS-R without stating a principled stopping point for how many transformer layers should remain trainable, a question the layer-wise interpretability literature discussed in Chapter 3 speaks to more directly.

No single study identified in this review combines all three elements. A dedicated Urdu dataset, a bidirectional cross-lingual evaluation design, and an XLS-R layer-weighting architecture have, to date, been developed separately rather than together, which is the specific combination this thesis brings together in Chapter 3.

Figure 2.2 illustrates the scale of the generalisation gap that motivates this combination, plotting the reported Equal Error Rate degradation across these and other studies when moving from within-language or within-corpus evaluation to cross-lingual or cross-domain conditions.

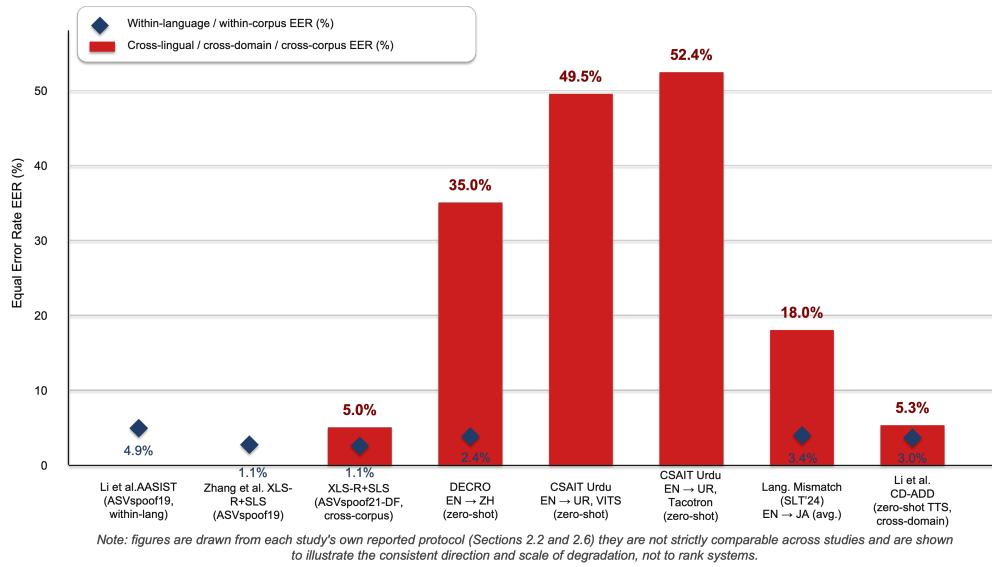


FIGURE 2.2: Reported EER degradation from within-language or within-corpus to cross-lingual or cross-domain evaluation conditions.

2.2 Foundational Benchmarks: The ASVspoof Lineage

ASVspoof is a series of international challenges, first launched in 2015 and now in its fifth edition, that provide standardised datasets, evaluation protocols, and metrics for benchmarking countermeasures against synthetic speech, voice conversion, and replay attacks. It is the single most widely used benchmark family in the anti-spoofing literature, and most of the field's infrastructure traces back to it. Liu et al. [8], reporting ASVspoof 2021 in IEEE/ACM Transactions on Audio, Speech, and

Language Processing, summarise 54 teams across logical access, physical access, and a newly introduced deepfake (DF) track targeting codec-processed speech. Countermeasures held up reasonably well against the LA task’s new transmission conditions but generalised poorly across the different source corpora bundled into the DF set, an early warning sign for the generalisation problems that dominate later cross-lingual work. The entire ASVspoof 2019/2021 pipeline is built from English, VCTK-derived speech only [1], meaning the field’s single most cited benchmark offers no direct evidence about cross-lingual behaviour.

On the architecture side, Jung et al. [3] introduced AASIST at ICASSP 2022, a graph-attention model that jointly reasons over spectral and temporal artefacts using a RawNet2-style [4] raw-waveform front end. AASIST cut roughly 20% off the prior state-of-the-art EER on ASVspoof 2019 LA, and a lightweight 85k-parameter variant, AASIST-L, stayed competitive with far larger systems. Complementary training techniques from the same period, RawBoost [22] (convolutive and additive waveform perturbation to simulate channel variability) and wav2vec 2.0 fine-tuning for anti-spoofing [23], have since become near-standard components of anti-spoofing training recipes.

The newest edition, ASVspoof 5 [24, 25], moved from lab-recorded to crowdsourced English speech from roughly 2,000 speakers and, for the first time, added seven adversarial attacks (Malafide- and Malacopula-style perturbations [26]) alongside more than twenty conventional TTS/VC systems. Fifty-three teams submitted systems. Adversarial and codec-processed conditions were, unsurprisingly, the hardest to generalise across. ASVspoof 5 still does not treat language diversity as a primary axis of difficulty.

2.3 The XLS-R Backbone and Layer-Wise Representation Structure

XLS-R is a large, publicly released speech encoder that learns general-purpose representations of audio directly from raw waveforms, without any labelled data, by

predicting masked portions of its own internal representation during pretraining. It is pretrained on 436,000 hours of unlabelled speech spanning 128 languages, making it a natural candidate backbone for a cross-lingual detection task. A parallel strand of research asks not which classifier to build on top of a self-supervised encoder such as this, but what the encoder itself already knows, and at what depth. This question, and how it shapes the fine-tuning strategy used in this thesis, is taken up directly in Chapter 3.

2.3.1 XLS-R and wav2vec 2.0

XLS-R [9], presented at Interspeech 2022 by Babu, Wang, Tjandra, and colleagues, is a wav2vec 2.0-style model [27] pretrained with self-supervised contrastive learning on 436,000 hours of speech across 128 languages, explicitly designed to yield cross-lingual representations. The English benchmark most commonly used to fine-tune and evaluate such models, ASVspoof 2019 LA, is documented in Wang et al. [1].

2.4 Other Self-Supervised Front-Ends for Spoof Detection

Zhang et al. [21], at ACM Multimedia 2024, pair XLS-R with a sensitive layer-selection (SLS) classifier that learns a weighting over the model’s 24 transformer layers rather than reading out only the final layer, reducing EER on ASVspoof 2019 from a prior state of the art of 4.07% to 1.15%, and reporting 4.98% EER on the harder ASVspoof 2021 DF set with a related XLS-128-based system. Cao et al. [28], in a 2026 Springer chapter, propose a dilated dual-branch classifier over XLS-R features that combines a multi-scale convolutional branch with a sensitive-layer branch, aiming to capture fine local artefacts and broader contextual cues simultaneously, reporting improved generalisation over the single-branch XLS-R+SLS-baseline. Guo et al. [29], at ICASSP 2024, instead pair WavLM [30]

with a multi-fusion attentive classifier that pools information across both time and layer using attentive statistics, reporting competitive results on ASVspoof 2019 and 2021.

None of these three studies tests its encoder on a genuinely unseen language. Multilingual pretraining is present in the XLS-R backbone but is never evaluated cross-lingually in any of them. Collectively they do converge on two recurring design choices. An end-to-end, fine-tuned self-supervised encoder over raw audio rather than a fixed feature extractor over spectrograms, and a learned weighting over multiple transformer layers rather than a single fixed layer.

2.5 Cross-Lingual and Multilingual Generalisation

The most directly relevant single study addressing cross-lingual transfer explicitly is Ba et al. [16], “Transferring Audio Deepfake Detection Capability across Languages”. The authors identify the English-only bias of existing countermeasure research as an unaddressed problem and construct DECRO, a paired deepfake cross-lingual evaluation dataset spanning English and Mandarin Chinese subsets of comparable duration (137 hours total). Using DECRO, they show a detector trained purely on English deepfakes suffers substantial performance degradation when tested on Chinese deepfakes, confirming that spoofing artefacts learned in one language’s feature space do not trivially transfer to another. They compare a naive joint-training baseline, straightforward fine-tuning on the target language, and an adversarial domain-adaptation (ADA) approach that explicitly trains the model to discard language-identifying information while retaining real-versus-fake discriminative information, finding the adversarial approach outperforms both simpler alternatives.

Chetia Phukan et al. [18] compare eight monolingual, multilingual, and speaker/emotion-oriented self-supervised pretrained models as front-ends for cross-lingual

spoof detection, evaluating on ASVspoof 2019, In-the-Wild, and DECRO. Representations from multilingual pretrained models consistently achieve the lowest EER across all three benchmarks, and a fusion framework, MiO (Merge into One), improves results further, supporting the idea that multilingual pretraining breadth, not just model scale, drives cross-lingual transfer. Nguyen et al. [31], at EMNLP 2025, complicate this picture by showing that both open-source and commercial anti-spoofing detectors are highly sensitive to linguistic properties of the input text, not just its acoustic content, introducing transcript-level adversarial perturbations (TAPAS) that achieve attack success rates exceeding 60% on several detector-voice pairs, with one commercial detector’s accuracy collapsing from 100% to 32% under minor linguistic perturbation.

At larger scale, Muller et al. [7] introduce MLAAD, the Multi-Language Audio Anti-Spoofing Dataset, at IJCNN 2024, spanning over 170 TTS systems and more than 1,000 hours of synthetic speech across more than 50 languages, generated by translating and resynthesising the multilingual M-AILABS speech corpus. Training on MLAAD outperformed training on comparable single-purpose datasets across eight cross-dataset test sets, and MLAAD and ASVspoof 2019 were found complementary rather than redundant, each outperforming the other on four of eight evaluations. Mawalim et al. [20], building a system for the 2025 SAFE Challenge, construct a 17-language corpus and a hybrid CNN-plus-SSL architecture, finding performance varies considerably by language and attack algorithm, and that naively combining heterogeneous data sources can degrade performance on individual subsets if class balance and generator diversity are not controlled.

Liu et al. [32], at IEEE SLT 2024, directly quantify the language-mismatch effect. Training four state-of-the-art architectures exclusively on English and testing on Japanese and other languages, they estimate language mismatch alone causes a relative EER degradation exceeding 15%, with partial recovery available through limited target-language fine-tuning. Marek et al. [19], in “Are Audio DeepFake Detection Models Polyglots?”, reach a compatible conclusion. A small amount of target-language data outperforms a larger volume of data drawn from several non-target languages combined.

2.6 Low-Resource Language Datasets and Detection: Urdu and Beyond

Central to the English–Urdu cross-lingual literature is Munir et al. [17], “Deepfake Defense. Constructing and Evaluating a Specialized Urdu Deepfake Audio Dataset”, published in Findings of the Association for Computational Linguistics. ACL 2024. This paper constructs the CSALT Urdu Deepfake Audio Dataset, generating spoofed Urdu speech using Tacotron and VITS TTS systems and pairing it with bona fide Urdu recordings drawn from Urdu news sources. It is, as far as this review could establish, the earliest peer-reviewed, dedicated English–Urdu deepfake resource in circulation, has already been reused as training or evaluation data by at least one independent follow-on study [33], and reports both automatic countermeasure results and a controlled human-perception study.

Two findings from Munir et al. [17] are especially significant. First, evaluating a pretrained AASIST-L countermeasure, trained on English ASVspoof data, directly on Urdu deepfakes produced EERs of 49.5% (VITS) and 52.4% (Tacotron), statistically indistinguishable from chance, direct empirical evidence that English-trained countermeasures do not transfer to Urdu without adaptation. Second, a human-perception study found ordinary listeners could distinguish genuine from synthetic Urdu speech only slightly better than chance (ROC AUC of 0.63), meaning roughly one in three fake Urdu audio samples were judged genuine, underscoring that the absence of automated Urdu-language countermeasures cannot be compensated for by human vigilance.

The CSALT Urdu dataset and its associated benchmark have begun to accumulate independent follow-on work. Owais et al. [33], in a 2026 IEEE Access paper, construct baseline, augmented, and extended configurations derived from the same underlying Urdu corpus and systematically benchmark convolutional (LCNN, CNN–LSTM with attention) and transformer-based (Whisper-variant) architectures over MFCC and LFCC front-ends, reporting accuracy, precision, recall, F1, and EER for each configuration.

This is, to this review’s knowledge, the most recent published benchmark specifically targeting Urdu deepfake detection. Azeemi et al. [34], drawing on the same CSALT-LUMS research group, propose a self-supervised dataset-pruning method for efficient anti-spoofing training at Interspeech 2023, indicating an active, connected research programme around low-resource countermeasure development at the dataset’s originating institution. Other low-resource-language efforts include the Indonesian Speech Anti-Spoofing System paper by Arief et al. [35], which documents the construction of Indonesian-language anti-spoofing dataset and baseline convolutional-network detectors, and a follow-on study by Prihasto et al. [36] that integrates RawBoost and SpecAugment-style data augmentation with a wav2vec 2.0 front-end specifically to improve generalisation on the same Indonesian dataset.

Both echo the CSALT Urdu effort’s core motivation that widely spoken but under-resourced languages currently have little to no dedicated deepfake countermeasure research despite hosting hundreds of millions of speakers. Across these unrelated research groups and languages, one pattern recurs without exception. Wherever a genuinely new, previously unstudied language is evaluated against an English-only-trained detector, near-chance performance is reported, arguably the single most consistent empirical finding in the entire cross-lingual literature reviewed here.

2.7 Class Imbalance and Training Dynamics

A separate, smaller strand of the literature examines how the training process itself is shaped by the class imbalance nearly every anti-spoofing corpus exhibits, ASVspoof 2019 LA included, where spoofed utterances outnumber bonafide ones by roughly nine to one in the raw training partition. Shim et al. [37], presented at the ASVspoof 5 Workshop (Interspeech 2024 satellite), examine this imbalance from the inside, analysing per-class training loss curves rather than only final error rates. They find that even after applying class weights, 0.9 to the bonafide class and 0.1 to the spoof class in their setup, a persistent gap remains between how quickly the two classes are learned, with the bonafide class showing signs of being

systematically harder to model well despite receiving the larger loss weight. They further flag that silence duration differs measurably between bonafide and spoofed recordings, and that a detector can learn to exploit this difference as a shortcut rather than learning genuine synthesis-artefact cues.

Related work has proposed weighted margin losses to address the same 1:9 imbalance in ASVspoof 2019 with class-specific loss weights, and focal-loss variants have been used for the same purpose in replay-attack detection, both consistent with the general prescription that class-imbalance handling has become a standard, if under-discussed, part of anti-spoofing training recipes.

2.8 Domain Shift, Codec Robustness and Input Representation

A generalisation axis distinct from language, but closely related in spirit, concerns codec compression, channel conditions, and unseen generators. Muller et al. [38], “Does Audio Deepfake Detection Generalize?”, Interspeech 2022, is among the most cited cautionary papers in the field and also introduced the widely used In-the-Wild dataset of celebrity and political-figure speech.

Training detectors on one of Fake-or-Real, ASVspoof, or In-the-Wild and testing on the other two, the authors found reported in-domain accuracy above 99% collapsing toward chance level under cross-corpus evaluation, with the likely culprit being detectors that key on dataset-specific recording conditions and channel artefacts rather than generator-intrinsic synthesis cues.

The ASVspoof 2021 DF-track analysis [8] reported a compatible pattern for codec-processed speech specifically, and Xie et al. [39], introducing the CodecFake dataset in a 2025 IEEE/ACM Transactions paper, show that detectors never exposed to modern neural audio codecs during training suffer pronounced extra degradation relative to detectors exposed only to older DSP-based codecs.

2.8.1 Raw Waveform Versus Spectrogram Input

A more specific design question in the field concerns whether to feed a detection front-end hand-computed spectral features or the raw waveform directly. Di Pierno et al. [5], in a 2025 preprint proposing RawNetLite, argue explicitly against spectrogram-based and handcrafted-feature pipelines (MFCC, CQCC) on the grounds that such representations can discard fine-grained temporal cues useful for spotting synthesis artefacts, reporting strong cross-dataset results, 99.7% F1 in-domain and up to 83.4% F1 under a challenging out-of-distribution setting, using an end-to-end architecture trained directly on waveforms.

This is consistent with earlier evidence from the AASIST line of work [3] and RawNet2 [4], both of which also operate on raw audio, and with the XLS-R+SLS system [21], which likewise treats the waveform as the model’s direct input rather than pre-extracting a spectrogram. No study identified in this review directly compares raw-waveform against spectrogram-fed variants of the same self-supervised encoder on a cross-lingual English–Urdu detection task specifically, leaving this a question the wider raw-versus-handcrafted-feature literature can inform but not settle for that particular language pair.

2.9 Surveys and Expert Assessments

Several substantial review articles synthesise the wider field and function, in this review, as a proxy for aggregated expert opinion. Wang et al. [40], “Deepfake Detection. A Comprehensive Survey from the Reliability Perspective”, published in ACM Computing Surveys, reframes deepfake detection literature across audio, image, and video around the question of reliability under distribution shift rather than raw accuracy, arguing the field has over-optimised for benchmark leaderboards at the expense of deployment-grade robustness. Cross-lingual and cross-domain transfer are treated as a central instance of this reliability gap. Dixit et al. [41], “Review of Audio Deepfake Detection Techniques. Issues and Prospects”

in Expert Systems, catalogues feature-based, deep-learning, and hybrid detection approaches, and explicitly flags the scarcity of non-English datasets and the near-total absence of low-resource-language benchmarks as one of the field’s most significant open issues.

A 2025 Springer Cybersecurity paper by Firc et al. [42] makes a methodological argument directly relevant to how the comparison tables in Sections 2.12 and 2.13 of this chapter should be read. Because different studies use different datasets, protocols, and even different definitions of equal error rate computation, direct numeric comparison across papers is unreliable, and the field needs a standardised evaluation framework analogous to what ASVspooof provides for English. The authors demonstrate their proposed framework by re-evaluating 40 state-of-the-art deepfake speech detectors under harmonised conditions. Nasr et al. [43] extend the scope beyond synthesis-based deepfakes to splicing and copy-move manipulation, identifying benchmark fragmentation and dataset scarcity outside English and Mandarin as complementary problems to purely detection-based countermeasures.

Two further contributions, Mubarak et al. [44], writing in IEEE Access, treating visual, audio, and text-based deepfakes as a single unified phenomenon across 255 screened articles, and Zhang et al. [45], a March 2025 Sensors article providing the most audio-specific and quantitatively grounded survey identified in this review, round out a consistent four-source expert consensus. Benchmark accuracy figures substantially overstate real-world and cross-lingual reliability. Non-English and especially low-resource-language coverage remains the field’s most conspicuous gap. And methodological standardisation across studies is currently insufficient to support strong comparative claims.

2.10 The Generative Side: Why Cross-Lingual Detection Keeps Getting Harder

The detection literature reviewed in Sections 2.2–2.9 does not exist in isolation. It is a continual response to advances on the generative side. Li et al. [6], at EMNLP

2024, motivate their cross-domain benchmark specifically by the rise of zero-shot TTS systems capable of cloning a speaker’s voice from a single reference utterance, a capability that removes the practical bottleneck, needing many minutes of a target speaker’s recorded voice, that limited earlier voice-cloning attacks. Through attack-augmented training, fine-tuned Wav2Vec2-large and Whisper-medium encoders achieve 4.1% and 6.5% EER respectively on their new cross-domain benchmark, and few-shot adaptation, exposing a trained detector to only a handful of examples from a new domain or generator, recovers most of the lost performance.

Because zero-shot cloning systems are commonly trained on large multilingual speech corpora to begin with, they inherit multilingual capability by default, meaning an attacker no longer needs a language-specific synthesis pipeline to target a new-language victim. This is precisely the scenario the CSALT Urdu dataset paper [17] and the Indonesian anti-spoofing paper [35] respond to using comparatively conventional, language-specific TTS systems (VITS and Tacotron), which likely represents a floor rather than a ceiling on the acoustic realism attackers can now achieve, since neither paper evaluates against more capable zero-shot cloning systems.

This generative trajectory has a direct implication for cross-lingual detection research broadly. The useful lifespan of any fixed cross-lingual dataset, DECRO [16], MLAAD [7], or CSALT Urdu [17], is inherently limited, tied as each is to specific TTS/VC systems that will eventually be superseded by newer architectures with different, potentially less detectable, artefact signatures. This is part of why MLAAD is maintained as a versioned, continually expanding dataset [7] rather than a single static release, and why ASVspoof 5 [24, 25] deliberately incorporated adversarial attack systems designed to actively defeat existing countermeasures.

2.11 Most Recent Advances (2025–2026)

Because the field is moving quickly, this section isolates the most recently published contributions identified in this review, prioritised specifically for their recency, so

the chapter reflects the current, not merely the historically canonical, state of the art approaches.

Ciobanu et al. [46] introduce XMAD-Bench, accepted to Findings of ACL. EACL 2026, a large-scale cross-domain multilingual audio deepfake benchmark comprising 668.8 hours of real and fake speech across seven languages (Arabic, English, German, Mandarin Chinese, Romanian, Russian, and Spanish), explicitly constructed so speakers, generative methods, and real-audio sources differ between training and test splits.

Their central finding, that reported accuracy levels near 99% on in-domain benchmarks do not reflect performance under simultaneous cross-domain and cross-lingual shift, reinforces the pattern documented for the English–Urdu pair in Section 2.6 and extends it to a substantially broader set of languages and generators, though notably still without Urdu coverage.

At the evaluation-methodology level, Kwok et al. [47], at Interspeech 2025, show that even bona fide (genuine) speech cross-testing, evaluating a detector’s behaviour on genuine audio drawn from a corpus different from its training distribution, reveals systematic weaknesses in current countermeasures, independent of any spoofing-artefact question at all.

Qin et al. [48], at APSIPA ASC 2025, propose a variational regularisation objective for end-to-end speech deepfake detection aimed at improving generalisation without requiring target-domain data, an approach conceptually adjacent to the adversarial domain-adaptation strategy of Ba et al. [16] but applied at the level of the training objective rather than an explicit language-identity discriminator.

Owais et al. [33], published in IEEE Access in 2026, represents the most recent dedicated Urdu deepfake detection benchmark identified in this review (Section 2.6). Taken together, the 2025–2026 literature confirms that none of the field’s most recent advances, larger multilingual benchmarks, improved cross-corpus evaluation protocols, or new regularisation objectives, has yet closed the cross-lingual and low-resource-language gap identified throughout this chapter.

2.12 Methodology Comparison

Table 2.1 compares the modelling methodology of studies that propose or benchmark a detection architecture. Dataset-only contributions, where no new architecture is proposed, are covered separately in Table 2.2.

TABLE 2.1: Comparison of model architectures, datasets, and reported results across principal studies reviewed.

Study	Model	Dataset	Cross-Lingual Setting	Result
Munir et al. [17] (2024)	AASIST-L	Urdu Deepfake Audio Dataset (self-created, PRUS corpus)	Explicit (Urdu, zero-shot)	49.5–52.4% EER (approximately chance-level performance)
Zhang, Wen & Hu [21] (2024)	XLS-R + SLS classifier	ASVspooF 2019 LA, ASVspooF 2021 DF, In-the-Wild	Implicit (SSL pretraining only)	1.15% EER (English only)
Ba et al. [16] (2023)	RawNet2, AASIST, RES-TSSDNet	ASVspooF 2019, DECRO (English/Chinese), FMFCC-A	Explicit (English ↔ Chinese)	11.92% EER after domain adaptation
Jung et al. [3] (2022)	RawNet2 + graph attention	ASVspooF 2019 LA	None (English only)	0.83% EER (AASIST-L: 0.99%)
Muller et al. [38] (2022)	Multiple architectures	ASVspooF 2019, In-the-Wild (self-collected)	None (cross-dataset only)	Lowest 1.2% EER on ASVspooF; rising to 33.9–37.4% EER on In-the-Wild
Owais et al. [33] (2026)	LCNN, CNN–LSTM+Attention, Whisper-tiny/base/small	Urdu Deepfake Audio Dataset (base-line/augmented/extended)	Explicit (Urdu, low-resource language)	Best: Whisper-small 0.50% EER (extended dataset) 38.90% EER Augmented dataset, worst: 42.20% EER (CNN+MFCC baseline)

Table 2.1 (Continued): Comparison of model architectures, datasets, and reported results across principal studies reviewed.

Study	Model	Dataset	Cross-Lingual Setting	Result
Xie et al. [39] (2025)	W2V2-AASIST + CAVG-based CM	CodecFake, ASVspoof 2019/2021/5, MLAAD, In-the-Wild	Explicit (English vs. Chinese source comparison)	Lowest EER on diverse tests; English-trained 0.616%, 2.770%, EER vs. Chinese-trained 7.024% EER
Firc et al. [42] (2024)	Temporal Convolutional Network (TCN) + multiple spectrograms	WaveFake, FakeAVCeleb, FoR, ASVspoof 2019 LA / 2021 DF, English/Czech	Explicit (English $\leftrightarrow$ Czech)	Same dataset: 2.13–23.37% EER (FoR, ASV) cross-dataset: 45.85–49.05% EER
Li et al. (2023)	HuBERT + ECAPA-TDNN	ASVspoof 2019/2021, FMFCC-A, FAD, THCHS-30	Explicit (English + Chinese)	2.89% EER; 1% EER with augmentation + α -FMS
Sharafat et al. (2025)	CNN-LSTM hybrid	ASVspoof 2021 LA eval + DEEP-VOICE (combined)	None (English only)	97.0% accuracy, 97.4% F1, EER 9–11%
Liu et al. [32] (2024)	SENet / Res2Net / AASIST	ASVspoof 2015/17/19/21, SingFake, FakeAVCeleb, Indonesian	Explicit (English $\rightarrow$ Japanese)	more than 15% relative EER degradation
Marek et al. [19] (2024)	W2V+AASIST, LFCC-based models, RawGAT-ST	ASVspoof 2019 (EN fine-tuning base), MLAAD v3 (23 languages), M-AILABS	Explicit (8 languages, cross-lingual FT)	EER varies by language, $\pm 8.91\%$ EER; Russian hardest at 15.74% EER

Table 2.1 (Continued): Comparison of model architectures, datasets, and reported results across principal studies reviewed.

Study	Model	Dataset	Cross-Lingual Setting	Result
Li et al. [6] (2024)	Wav2Vec2-large / Whisper-medium	CD-ADD (5 zero-shot TTS models), ASVspoof 2019	Cross-domain (zero-shot TTS/VC)	4.1–6.5% EER (adaptation recovers)
Di Pierno et al. [5] (2025)	RawNetLite (raw-waveform)	FakeOrReal, AVSpooF2021, CodecFake	None (cross-dataset only)	FoR 0.25% EER, AVSpooF2021 17.34%, CodecFake 50.68%; cross-domain 36.88% EER

The comparison reveals three main trends. That highlights the key differences and similarities among the reviewed studies.

First, there is a clear historical shift from architecture-driven approaches evaluated on a single language (AASIST [3]) toward approaches that pair a large multilingual self-supervised encoder with comparatively lightweight or dual-branch classification head [21, 28, 29].

Second, only a minority of studies, Ba et al. [16], Chetia Phukan et al. [18], Liu et al. [32], Marek et al. [19], Mawalim et al. [20], and the CSALT Urdu study [17] together with its 2026 follow-on [33], actually evaluate cross-lingual transfer directly, rather than relying on the implicit assumption that multilingual pretraining alone guarantees cross-lingual detection performance.

Third, nearly every recent cross-lingual-aware study pairs a self-supervised or CNN-based encoder with a comparatively simple classification stage, suggesting the field’s remaining performance gap on cross-lingual transfer is currently attributed more to representation-level generalisation than to the design of the final classifier.

2.13 Datasets Used: Comparative Table

TABLE 2.2: Datasets referenced across the reviewed literature, along with languages details and publicly availability.

Dataset/ Languages	Introducing Study / Venue	Publicly Available?	Scale	Notes
ASVspoof 2019 LA (English only)	Wang et al. [1], summarised in Liu et al. [8]	Yes (Zenodo, open download)	~121,000 utterances, 19 attack types (A01–A19)	Field’s flagship benchmark, widely used as an English fine-tuning source
CSALT Urdu Deepfake Audio Dataset (Urdu)	Munir et al. [17], ACL Findings	Yes (CC BY-NC 4.0, non-commercial)	~3,398 bona fide files (42.9 hours) plus Tacotron-VITS spoofed sets	First dedicated Urdu deepfake dataset reused and extended by [33]
DECRO (English, Mandarin Chinese)	Ba et al. [16], ACM WWW	Yes (Zenodo, CC BY 4.0)	137 hours total audio	Dedicated cross lingual deepfake evaluation dataset
ASVspoof 2021 (LA/PA/DF) (English only)	Liu et al. [8], IEEE/ACM TASLP	Yes (Zenodo, open download)	54 participating teams’ submissions evaluated	DF track shown to generalise poorly across source sub-datasets
ASVspoof 5 (English, crowdsourced)	Wang et al. [24, 25]	Yes (Zenodo / Hugging Face, usage terms apply)	~2,000 speakers, 32 attack systems incl. 7 adversarial [26]; 11 codec conditions	Most recent (2024–2026) English benchmark, first adversarial attack tracks
MLAAD (v9–v10) (51–54 languages)	Muller et al. [7], IJCNN	Yes (public release, versioned)	~687–1,003 hours, 140–175 TTS models	Largest multilingual anti spoofing corpus from M-AILABS
In-the-Wild (predominantly English)	Muller et al. [38], Interspeech	Yes (public download)	Real-world celebrity/ public audio, bona fide and spoofed	Tests cross-corpus, real-world generalisation, not language diverse

Table 2.2 (Continued): Datasets referenced across the reviewed literature.

Dataset/ Languages	Introducing Study / Venue	Publicly Available?	Scale	Notes
Fake-or-Real (FoR) (English)	Referenced in Muller et al. [38] and MLAAD baselines [7]	Yes (public download)	Moderate-scale, multiple TTS sources	One of three classic English corpora used in cross-corpus generalisation studies
WaveFake (English, Japanese)	Frank & Schonherr [49]	Yes (Zenodo, open download)	~196 hours synthetic audio across multiple neural vocoders	Early (2021) bilingual spoofing dataset; still used as a secondary benchmark
XMAD-Bench (Arabic, English, German, Mandarin, Romanian, Russian, Spanish no Urdu subset)	Ciobanu et al. [46], Findings ACL: EACL 2026	Yes (publicly released)	668.8 hours of real and deep-fake speech	Most recent (2026) cross-domain multilingual benchmark; disjoint train/test speakers and generators

The dataset landscape summarised in Table 2.2 makes visible a structural imbalance. Of the eleven listed datasets, four (DECRO, MLAAD, WaveFake, XMAD-Bench) are explicitly cross-lingual or multilingual by design, and only two (CSALT Urdu and the Indonesian dataset) target a single low-resource language outside English or Mandarin. ASVspoof, by far the most cited resource family and now extended through its fifth edition, remains monolingual, meaning a very large share of the field’s reported progress is still measured almost exclusively on English speech. XMAD-Bench [46] partially closes this gap at the multilingual end, but no dataset identified in this review yet pairs the scale of MLAAD or XMAD-Bench with dedicated South Asian language coverage, leaving the CSALT Urdu dataset as the most prominent dedicated South Asian-language resource identified in this review.

2.14 Summary of Research Gaps

Drawing the threads of Sections 2.1 through 2.11 together, four recurring gaps emerge from the literature surveyed in this chapter.

- i. English-only benchmark dominance. The field’s flagship resource, ASVspoof, remains monolingual through its fifth edition [24, 25], and the great majority of self-supervised front-end papers surveyed in Sections 2.3 and 2.4 report English-only results despite using multilingually pretrained encoders such as XLS-R. This progress has not extended to Urdu.
- ii. Urdu remains almost entirely unaddressed outside two papers, Munir et al. [17] and its direct follow-on Owais et al. [33], neither of which combines Urdu training data with a larger majority-language pool to test recovery, and neither of which uses a raw-waveform self-supervised encoder trained end-to-end.
- iii. Bidirectional cross-lingual transfer, tested by Ba et al. for English–Mandarin [16] and by Liu et al. for English-to-Japanese [32], has not previously been reported for the English–Urdu pair in either direction using a single consistent architecture.

These gaps, taken together, motivate the methodology described in Chapter 3.

Chapter 3

Proposed Methodology

Voice-based deepfakes have moved from a research curiosity to a practical security concern within the space of only a few years. Speech synthesis systems that once produced obviously robotic output can now generate utterances that are difficult for an untrained listener to distinguish from genuine human speech. The consequences of this progress are not hypothetical. Fraudulent transactions authorised through cloned executive voices, fabricated statements attributed to public figures, and impersonation attempts targeting call-centre verification systems have all been documented in recent years. Detecting such content reliably has therefore become an active and urgent area of speech-processing research.

The overwhelming majority of published detection systems, however, are built and evaluated almost exclusively on English-language corpora. This creates a blind spot with direct real-world consequences for multilingual regions such as South Asia, where Urdu, Punjabi, Pashto, and related languages are spoken by hundreds of millions of people. A detector trained only on English speech offers no assurance that it will behave sensibly when confronted with a language it has never encountered. And, as this chapter will demonstrate empirically, that assurance would in fact be misplaced. This research was undertaken to examine precisely this question. To what extent does a deepfake detector trained on one language transfer to a linguistically distant second language, and can that gap be narrowed through deliberate bi-lingual training?

To answer this question in a controlled and reproducible manner, the study centres on two languages chosen for their typological distance from one another. English is represented through the ASVspoof 2019 Logical Access corpus, the most widely cited benchmark in the anti-spoofing literature. Urdu is represented through the CSALT Urdu Deepfake Detection Dataset released at ACL 2024. The pairing of these two corpora allows the research to probe cross-lingual generalisation.

Rather than relying on hand-engineered acoustic features such as Mel-frequency cepstral coefficients or spectrogram representations, the dominant approach in earlier anti-spoofing work, this study adopts XLS-R, a cross-lingual self-supervised speech model pretrained on 436,000 hours of unlabelled audio spanning 128 languages, as the acoustic backbone. Because Urdu is among the languages present in XLS-R’s pretraining corpus, the model arrives already equipped with some degree of latent Urdu phonetic knowledge before any task-specific fine-tuning takes place, which makes it a particularly well-suited candidate for a cross-lingual investigation of this kind. Figure 3.1 illustrates the resulting end-to-end pipeline.

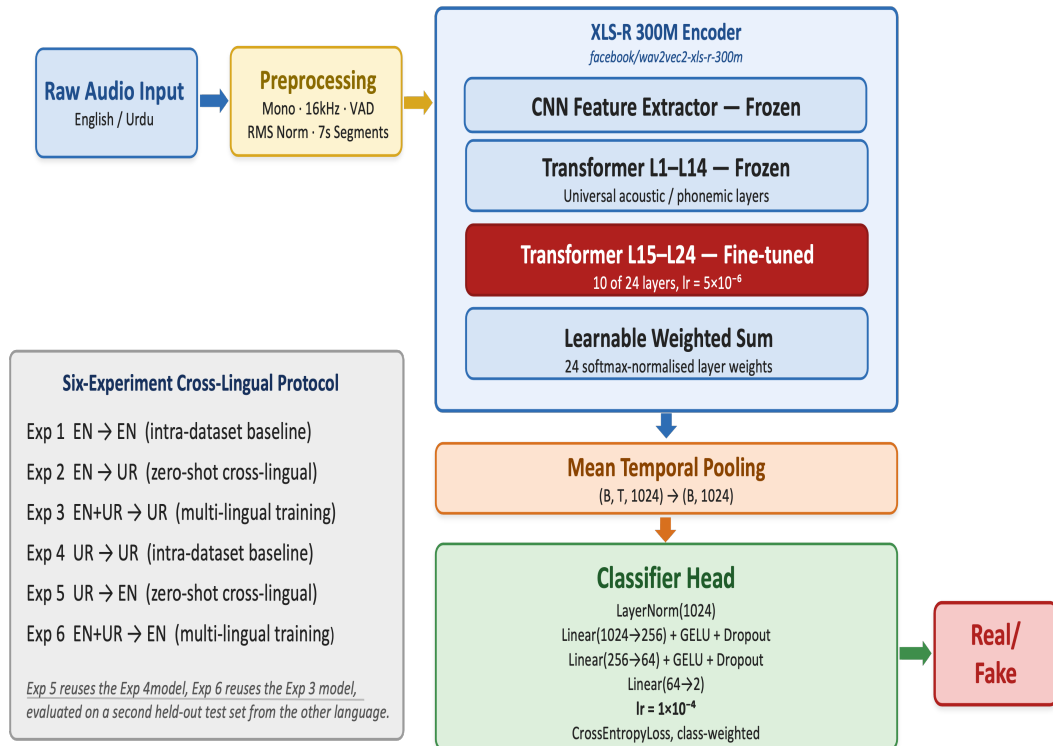


FIGURE 3.1: End-to-end architecture of the proposed cross-lingual deepfake voice detection system, from raw waveform input through the XLS-R encoder to the final real or fake decision.

3.1 Model Architecture: XLS-R 300M

At the centre of the detection system sits XLS-R [9], a cross-lingual variant of the wav2vec 2.0 self-supervised speech representation framework [27]. This section explains the model’s pretraining mechanism, its cross-lingual transfer characteristics, the way its internal representations are formed, and the fine-tuning strategy adopted in this study.

3.1.1 How XLS-R Works: A Multilingual Speech Recognition Model

XLS-R takes the wav2vec 2.0 design and scales it up so that it can handle many languages together instead of just one. To pretrain it, researchers feed the model very large amounts of unlabeled audio collected from dozens of languages, using a self-supervised approach. The raw waveform first passes through a convolutional feature encoder, which extracts low-level acoustic details from the sound. These extracted features then move into a Transformer network, which learns relationships across longer stretches of speech. During pretraining, portions of the audio are hidden from the model, and it has to work out which of several candidate latent speech units belongs in each hidden spot, using a contrastive learning task. Because this training signal comes from the audio itself and not from language-specific transcripts, the resulting representations are not locked to any one language.

3.1.2 Cross-Lingual Transfer

A major strength of XLS-R over single-language wav2vec 2.0 models is how well it carries knowledge from one language to another. Even with only a small labeled dataset for a low-resource language, fine-tuning XLS-R on that limited data can still produce large gains, because the model has already picked up general speech patterns such as phonetic structure, rhythm, and articulation during multilingual pretraining. Even so, XLS-R has clear limits. It needs a large and varied

pretraining dataset, otherwise it risks leaning too heavily toward whichever languages dominate the data. Its accuracy can also drop for languages with phonetic systems very different from what it saw in training, or when the recordings come from unusual or noisy conditions.

3.1.3 How the Model Extracts Features and Represents Language

The diagram 3.2 below breaks the process into four stages, moving from raw sound to a shared multilingual space of speech units.

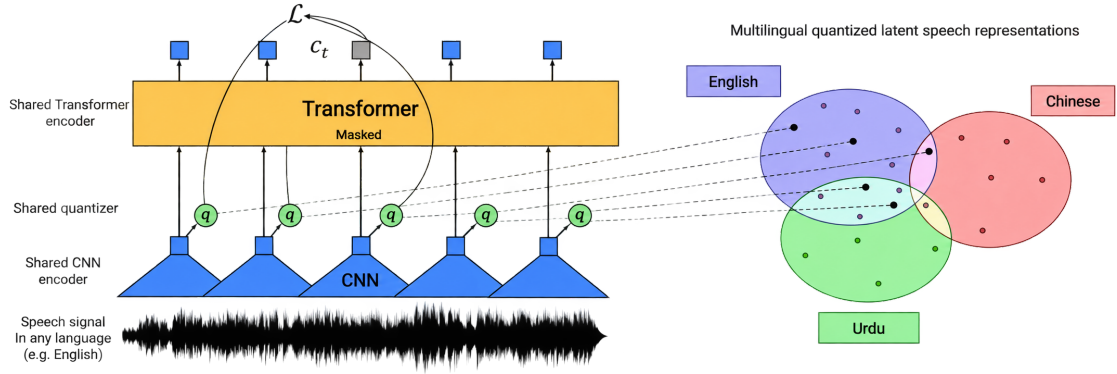


FIGURE 3.2: Feature extraction and multilingual representation pipeline in XLS-R.

3.1.3.1 Raw Audio to Local Features (CNN Encoder)

The process starts with the raw speech waveform, shown at the bottom of the diagram. A stack of convolutional layers, labeled the shared CNN encoder, slides across this waveform and turns it into a sequence of short feature vectors, marked as z_t in the figure. Each vector describes the local acoustic character of a small slice of audio, capturing details like pitch, energy, and formant shape.

At this stage, the model does not yet know which words are being spoken or which language is in use. The encoder is called shared because the same convolutional weights process audio from every language in the training set. There is no separate encoder built for each language.

3.1.3.2 Turning Sound into Latent Speech Units (Quantizer)

Next, the continuous vectors from the CNN encoder pass through a quantizer, marked as q in the diagram. This step maps each vector onto the closest entry in a shared codebook of discrete latent speech units. This shared codebook is the main reason XLS-R can work across languages. Instead of learning a separate codebook for every language, it learns a single inventory of sound units. Because many languages produce physically similar sounds, an Urdu speech sound and an English speech sound that are pronounced in a comparable way can end up mapped to the same or a nearby code. This is exactly what the right side of the diagram shows. In the multilingual quantized latent speech representations space, the circles for English, Chinese, and Urdu overlap in places. The black dots mark latent units shared across languages, while the dots sitting in the non-overlapping parts of each circle represent sounds that are more specific to that one language.

3.1.3.3 Learning Context with the Transformer and Contrastive Loss

Some of the z_t vectors are masked before they reach the Transformer encoder, shown as the masked positions in the middle of the diagram. The Transformer produces an output c_t at each masked position, and the model must pick out the correct quantized unit for that position from a small set of candidates, which include the true unit plus several distractors. This comparison is the contrastive loss shown at the top of the diagram as L . To succeed at this task across dozens of languages at the same time, the Transformer is pushed to build genuinely useful representations of phonetics, rhythm, and articulation, since those are the cues that let it separate the correct unit from the distractors no matter which language is playing.

3.1.3.4 How Does the Model Identify a Language

It helps to be precise here. During pretraining, XLS-R is never told outright that a clip is in Urdu or in English. There is no language-label classification step built

into this stage. What the model builds instead is a language-agnostic map of speech sounds, where similar sounds from different languages sit close together and more distinctive sounds stay apart, as pictured in the overlapping circles. Explicit language identification is a separate downstream task.

A team can take these pretrained latent representations and fine-tune a classifier on top of them using real language labels. The pretraining stage shown in the diagram is what makes that later step so effective, because the model already understands the general shape of speech, so it only takes a modest amount of labeled data to teach it to tell specific languages apart.

The remainder of this chapter documents the dataset acquisition and characterisation, audio preprocessing and segmentation, the detailed fine-tuning and classification components of the detection model, the treatment of class imbalance, and the metrics used to quantify performance. The training configuration and the six-experiment evaluation protocol used to produce the results are described at the start of Chapter 4, immediately before those results are presented.

3.2 Dataset Collection and Characterisation

Two corpora were selected for this research, one per target language, on the basis of public availability, peer-reviewed provenance, and sufficient scale to support meaningful fine-tuning of a 300-million-parameter model. Table 3.1 summarises the key characteristics of both corpora, before and after preprocessing, together with the combined pool used for the bi-lingual training experiments described later in this chapter.

3.2.1 ASVspoof 2019 Logical Access Partition

The Logical Access partition of the ASVspoof 2019 challenge corpus [1] supplies the English portion of this study. It contains genuine utterances recorded from 107 speakers drawn from the VCTK corpus, alongside spoofed counterparts produced

by nineteen distinct text-to-speech and voice-conversion systems, catalogued under the labels A01 through A19. Of these, six algorithms (A01–A06) are represented in the official training partition, while the remaining thirteen appear only in the development and evaluation partitions, a design choice intended to test generalisation to unseen synthesis methods. Both the official training and development partitions were combined for this study, since together they provide roughly fifty thousand labelled utterances and their aggregate class balance still permits an 80/20 stratified re-split without contaminating any downstream evaluation.

A pronounced class imbalance is immediately apparent in the raw data. Bonafide utterances number only 5,128 against 45,096 spoofed utterances, an imbalance ratio just under nine to one. Left uncorrected, this imbalance would allow a classifier to achieve superficially high accuracy simply by predicting the majority class for every input. A failure mode observed directly during early experimentation, where a naively trained model produced a confusion matrix with zero true positives on the bonafide class despite reporting a headline accuracy above 70%. Section 3.4 describes in detail how this problem was addressed through a combination of moderate random undersampling of the spoof class, reducing the ratio to a more tractable three to one while retaining every bonafide recording, and class-weighted loss applied to the resulting training pool to correct for the residual imbalance.

3.2.2 CSALT Urdu Deepfake Detection Dataset

The Urdu portion of the study draws on the CSALT Urdu Deepfake Detection Dataset [17], released by the Computational Semantics and Analysis of Language Technology laboratory at LUMS and presented at the 2024 Annual Meeting of the Association for Computational Linguistics. At the time this research was conducted, it represented the only publicly available, peer-reviewed Urdu-language deepfake audio corpus of comparable scale, making it the natural, and in practice the only realistic, choice for the Urdu arm of the study.

The corpus comprises 3,398 bonafide recordings contributed by seventeen native Urdu speakers, deliberately chosen to span a broad range of phonemic and prosodic

contexts, together with 3,396 synthetic counterparts generated using two neural text-to-speech systems. Tacotron 2 paired with a WaveGlow vocoder, and VITS, an end-to-end variational architecture known for producing highly natural prosody. The corpus is distributed as individual WAV files organised by speaker and recording part rather than through a single pre-built label column, so the bonafide or spoof status of each file was reconstructed directly from the underlying Hugging Face dataset metadata.

Duration analysis of the Urdu corpus revealed a substantially wider and longer distribution than its English counterpart. A mean of 8.47 seconds, a standard deviation of 3.48 seconds, and a maximum observed duration of 40.07 seconds. Recordings beyond the ninety-fifth percentile (15.0 seconds) were treated as outliers and excluded prior to segmentation, leaving 6,451 usable recordings while discarding fewer than three hundred and fifty atypically long files.

TABLE 3.1: Summary statistics for both corpora before and after preprocessing, together with the combined pool used for bi-lingual training experiments.

Property	ASVspoof 2019 LA (English)	CSALT Urdu Deepfake	Combined
Language	English	Urdu	English + Urdu
Original recordings	50,424	6,794	57,218
Bonafide (real)	5,128	3,398	8,526
Spoofed (fake)	45,096	3,396	48,492
Retained after filtering	20,785	6,451	27,236
Total 7s segments	20,785	12,593	33,378
Training segments	16,638 (80%)	10,073 (80%)	26,711 (80%)
Test segments	4,147 (20%)	2,520 (20%)	6,667 (20%)
Mean recording length	3.48 s	8.47 s	5.97 s
Maximum recording length	11.39 s	40.07 s	25.73 s
Segmentation expansion factor	$\times 1.01$	$\times 1.85$	$\times 1.22$

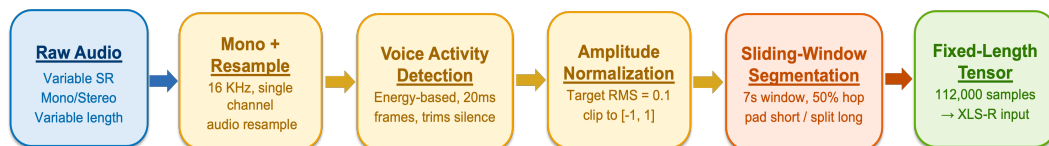
Table 3.1 (Continued): Summary statistics for both corpora.

Property	ASVspoof 2019 LA (English)	CSALT Urdu Deepfake	Combined
Spoof-to-bonafide ratio	3.05 : 1	1.17 : 1	2.05 : 1
Total audio duration	40.3 hrs	24.6 hrs	64.9 hrs
Sample rate	16,000 Hz	16,000 Hz	16,000 Hz
Audio format	FLAC	WAV → FLAC	FLAC

As Table 3.1 shows, the Urdu corpus is considerably smaller than its English counterpart in terms of original recordings, but the sliding-window segmentation strategy applied to its longer utterances partially offsets this gap, leaving the two languages with a more comparable number of training segments for the bi-lingual experiments described later in this chapter.

3.3 Audio Preprocessing and Segmentation

The two corpora differ in sample rate consistency, channel configuration, file format, and, most significantly, duration distribution, a uniform preprocessing pipeline was applied to every recording from both languages before any feature extraction took place. Applying identical processing to both languages was considered essential. Any performance difference observed later between English and Urdu should be attributable to genuine linguistic or acoustic factors, not to inconsistencies introduced during data preparation. Figure 3.3 summarises the five stages of this pipeline.



Identical pipeline applied to English (ASVspoof 2019 LA) and Urdu (CSALT) audio — no language-specific handcrafted filters are used

Files ≤7s are zero-padded to one segment; files >7s generate multiple overlapping segments, all inheriting the parent label

FIGURE 3.3: The audio preprocessing and segmentation pipeline.

3.3.1 Channel Normalisation and Resampling

Every recording was first collapsed to a single channel by averaging across available channels, and then resampled to 16,000 Hz using torchaudio’s sinc-interpolation resampler where the source sample rate differed. This step matters more than it might initially appear. XLS-R’s convolutional feature extractor contains fixed-stride convolutions whose receptive fields were calibrated during pretraining against audio sampled at exactly this rate, and feeding audio at a mismatched rate would silently misalign these receptive fields relative to what the model had learned to expect.

3.3.2 Voice Activity Detection

An energy-based voice activity detector, operating over 20-millisecond frames, was applied to trim leading and trailing silence from each recording. This addition was made after observing that fixed-length padding without silence removal was diluting the proportion of genuine speech content within each processed segment, particularly for the shorter ASVspoof clips where several hundred milliseconds of silence at either end of a four-second recording represented a meaningful fraction of the total signal. A conservative fallback was built into the detector. If the estimated active region fell below two seconds, the original untrimmed waveform was retained rather than risk discarding a genuinely quiet recording.

3.3.3 Amplitude Normalisation

Following silence removal, each waveform was rescaled to a target root-mean-square amplitude of 0.1 and clipped to the range $[-1, 1]$. This step was included to prevent the model from learning volume level as an incidental, and ultimately unreliable, cue for distinguishing genuine from synthetic speech. Text-to-speech systems typically normalise their output to a consistent loudness during generation, whereas naturally recorded bonafide speech carries whatever microphone gain

and speaker distance happened to be present at capture time. Without amplitude normalisation, a model could conceivably achieve artificially inflated performance on held-out data drawn from the same recording conditions while failing entirely on audio captured under different conditions.

3.3.4 Sliding-Window Segmentation

A fixed segment length of seven seconds (112,000 samples at 16 kHz) was adopted for both languages after examining the empirical duration distributions of each corpus. For English, this threshold captures 98.3% of recordings without any truncation. The mean ASVspoof utterance runs to only 3.48 seconds, so the great majority of English segments consist of genuine signal followed by a short zero-padded tail. For Urdu, whose mean duration of 8.47 seconds substantially exceeds this window, a sliding-window strategy with 50% overlap (a 3.5-second hop) was applied instead of simple truncation.

Recordings shorter than seven seconds were zero-padded into a single segment exactly as with English, while longer recordings were split into multiple overlapping seven-second windows, each inheriting the bonafide/spoof label of its parent recording. This design decision served two purposes simultaneously.

First, it avoided discarding the latter portion of longer Urdu recordings, where synthesis artefacts are just as likely to be present as in the earlier portion. An outright truncation strategy would have silently thrown away up to thirty seconds of signal from the longest files.

Second, it had the welcome side effect of substantially enlarging the Urdu training pool. The 6,451 retained Urdu recordings expanded to 12,593 individual training segments, a 1.85-fold increase, compensating in part for the smaller scale of the Urdu corpus relative to its English counterpart. The equivalent expansion factor for English was a negligible 1.01, confirming that the sliding window had no meaningful effect on the already-short ASVspoof recordings and that the observed

Urdu expansion was a genuine consequence of the corpus’s longer natural utterance length rather than an artefact of the segmentation logic itself.

3.3.5 Why the Uppermost Ten Layers Were Selected for Fine-Tuning

Guided by the layer-wise interpretation above, this study freezes the convolutional feature extractor together with the lower fourteen Transformer layers (L1 through L14), and fine-tunes only the uppermost ten layers (L15 through L24) alongside the newly initialised classifier head. Three considerations converged on this specific split point rather than a shallower or deeper one.

The first consideration is the risk of catastrophic forgetting. XLS-R’s 300 million parameters were shaped by 436,000 hours of pretraining audio. The combined English and Urdu training pool used here totals under 65 hours. Allowing gradient updates from a dataset several orders of magnitude smaller than the pretraining corpus to propagate freely through the entire network risks overwriting broadly useful representations with narrow, dataset-specific patterns. A well-documented failure mode when fine-tuning large pretrained speech and language models on comparatively small downstream datasets, consistent with prior investigations of which self-supervised front-end layers best serve spoofing countermeasures [26] and with wav2vec 2.0-based fine-tuning strategies for anti-spoofing more specifically [23]. Restricting updates to a bounded subset of layers acts as an implicit regulariser against this risk.

The second consideration is that a shallower cut-off, fine-tuning only the final five or six layers, as explored in earlier iterations of this work, was found empirically to leave the model with insufficient capacity to adapt convincingly to the detection task, particularly for the smaller and more attack-limited Urdu corpus. Extending the trainable region downward to layer fifteen (ten layers in total, representing a little under half of the Transformer stack) provided noticeably more headroom for the model to reorganise its higher-level representations around synthesis-artefact

detection, without reaching so far down as to disturb the phoneme- and speaker-level representations that the lower layers are understood to carry.

The third consideration was practical rather than theoretical. Fine-tuning ten layers rather than the full twenty-four kept the number of trainable parameters, and consequently the memory footprint and training time per epoch, within a range that permitted thirty full training epochs per experiment across six separate experimental configurations on the available compute budget.

A small, conservative learning rate of 5×10^{-6} was applied specifically to these fine-tuned backbone layers, an order of magnitude smaller than the 1×10^{-4} rate applied to the freshly initialised classifier head, reflecting the fact that the backbone weights already encode useful structure and require only gentle adjustment, whereas the classifier head begins from random initialisation and must learn its function largely from scratch.

3.3.6 Learnable Weighted Layer Aggregation

Rather than reading out a representation from a single, arbitrarily chosen Transformer layer, the twenty-four hidden-state outputs of the encoder are combined through a learnable weighted sum, as formalised in Equation 3.1.

$$\alpha_l = \text{softmax}(\theta)_l = \frac{\exp(\theta_l)}{\sum_{j=1}^{24} \exp(\theta_j)}, \quad h = \sum_{l=1}^{24} \alpha_l H_l \quad (3.1)$$

where θ is the vector of 24 learnable scalar weights, H_l is the hidden state output of transformer layer l , and h is the resulting weighted combination fed to the temporal pooling stage.

A vector of twenty-four scalar weights, initialised uniformly and passed through a softmax to guarantee they sum to one, is optimised jointly with the rest of the network during training. This allows the model to discover for itself which combination of representational depths is most informative for the detection task, rather than committing in advance to the assumption that the final layer alone

is optimal. An assumption that recent self-supervised speech literature has repeatedly shown to be unreliable across different downstream tasks. The overall training objective combines this weighting with the class-weighted cross-entropy loss defined in Equation 3.2

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N w_{y_i} \log(p(y_i | x_i)) \quad (3.2)$$

where x_i is the i -th input segment, y_i its ground-truth label, $p(y_i | x_i)$ the softmax probability assigned by the classifier head to the correct class, and w_{y_i} the class weight from Equation 3.3 corresponding to label y_i .

3.3.7 Temporal Pooling and Classification Head

The weighted combination of hidden states, still varying in length according to the duration of the input segment, is reduced to a single fixed-length utterance representation of 1,024 dimensions through mean pooling across the time axis. This utterance-level vector is then passed through a compact classification head consisting of a LayerNorm operation, followed by two hidden linear layers (1,024 to 256, then 256 to 64 dimensions) each followed by a GELU non-linearity and a dropout rate of 0.3%, and a final linear projection to two output logits corresponding to the bonafide and spoof classes. Dropout at this comparatively aggressive rate was retained throughout training as a regularisation measure appropriate to a classifier head trained on datasets that, while substantial by anti-spoofing standards, remain modest by the standards of the pretrained encoder feeding into it.

3.4 Class Imbalance and Role of Class Weighting

Class imbalance emerged as one of the most consequential practical issues encountered during this research, and the strategy adopted to address it is documented here in some detail because of the direct effect it has on the reliability of the reported results.

3.4.1 Why Class Weights Were Necessary

As noted in Section 3.2, the raw English corpus exhibits a spoof-to-bonafide ratio of roughly nine to one. Given the scale of this imbalance, and the risk that class weighting alone applied to the full, severely imbalanced pool could produce unstable gradient updates dominated by a small number of heavily-weighted bonafide examples, the spoof class was first randomly subsampled at a fixed three-to-one ratio relative to the bonafide class, using a fixed random seed for reproducibility. Every bonafide recording was retained, the spoof class was reduced from 45,096 to 15,384 recordings, discarding roughly two-thirds of the majority class while preserving substantially more spoof-class diversity than a full one-to-one balancing would have allowed. After preprocessing, segmentation, and the final train/test split, this yields a training-pool ratio of 3.05 to 1, close to but not identical to the nominal three-to-one target, the small discrepancy arising from files affected by voice activity detection and the stratified train/test partitioning.

A classifier trained under standard, unweighted cross-entropy loss on this still-imbalanced pool retains a readily available shortcut. Since the spoof class continues to outnumber the bonafide class more than three to one, a model that learns to predict spoof unconditionally, ignoring the acoustic content of the input entirely, will already achieve a substantial fraction of theoretically achievable accuracy without learning anything genuinely diagnostic, a bias in per-class training dynamics documented directly in the anti-spoofing literature [37]. This is precisely the failure mode observed during early experimentation on an unweighted configuration, where the resulting confusion matrix showed the model classifying every single bonafide test sample as spoof, despite an accuracy figure that superficially appeared respectable. Class weighting addresses this residual imbalance by modifying the loss function itself, applied on top of the undersampled training pool rather than as a substitute for it. Under a weighted cross-entropy formulation, a misclassification of the minority (bonafide) class is penalised more heavily than an equivalent misclassification of the majority (spoof) class, in direct proportion to the imbalance ratio observed in the training set. This has the effect of making

the two classes equally influential on the gradient signal during optimisation, regardless of how unevenly they remain represented numerically. The model is no longer rewarded for ignoring the minority class, because doing so now carries a proportionally larger loss penalty.

3.4.2 How the Weights Were Computed

For each training configuration, the weight applied to the bonafide class was calculated as the ratio of spoof to bonafide sample counts within that specific training partition, with the spoof class fixed at a baseline weight of 1.0. For the English-only training pool (12,530 spoof against 4,108 bonafide segments after the final train/test split), this yields a weight of 3.05 applied to the bonafide class. For the Urdu-only training pool (5,430 spoof against 4,643 bonafide segments), the corresponding ratio is a much milder 1.17, reflecting the fact that the Urdu corpus, having been constructed with an explicit one-to-one design between real and synthetic speakers, was considerably better balanced from the outset than its English counterpart. For the combined English-plus-Urdu training pool used in Experiments 3 and 6, the aggregate ratio of 17,960 spoof to 8,751 bonafide segments produces an intermediate weight of 2.05.

$$w_1 = \frac{N_0}{N_1}, \quad w_0 = 1 \quad (3.3)$$

where N_0 and N_1 are the number of spoof (label 0) and bonafide (label 1) segments respectively in the training partition, and w_0 , w_1 are the corresponding weights.

TABLE 3.2: Class weights applied to the CrossEntropyLoss function.

Training Configuration	Bonafide Spoof	Applied Weight	Used In
English only	1 : 3.05	[1.0, 3.05]	Experiments 1, 2
Urdu only	1 : 1.17	[1.0, 1.17]	Experiments 4, 5
English + Urdu combined	1 : 2.05	[1.0, 2.05]	Experiments 3, 6

Table 3.2 summarises the resulting bonafide-class weight for each of the three training configurations, alongside the underlying spoof and bonafide segment counts from which each ratio was derived.

3.4.3 Observed Impact of Class Weighting

The practical effect of this weighting scheme was substantial and immediately visible once introduced. Where the unweighted configuration collapsed to predicting the majority class throughout, the weighted configuration produced confusion matrices with meaningful numbers of both true positives and true negatives from the earliest training epochs onward, and the reported Equal Error Rate, a threshold-independent measure inherently sensitive to both classes, dropped from a degenerate, uninformative figure to the competitive values reported in Chapter 4. It is worth emphasising that class weighting adjusts the training signal only. It has no bearing on the composition of the held-out test set, which remains a faithful, unweighted sample of the true class distribution throughout, ensuring that reported test-time metrics reflect performance on realistic, naturally imbalanced data rather than an artificially rebalanced evaluation set.

3.5 Evaluation Metrics

Performance is reported using a combination of metrics standard to the anti-spoofing literature together with general classification metrics intended to make the results interpretable to readers without specialist anti-spoofing background.

3.5.1 Equal Error Rate

The Equal Error Rate, or EER, is the primary metric adopted throughout this study, consistent with its status as the standard reporting metric across the ASVspoof challenge series and the wider anti-spoofing literature. It is defined

as the point at which the false rejection rate, the proportion of genuine utterances incorrectly classified as spoofed, equals the false acceptance rate, the proportion of spoofed utterances incorrectly accepted as genuine, as the decision threshold is swept across its full range. Because EER is threshold-independent by construction, it remains meaningful for cross-lingual comparison even where the underlying score distributions differ between training and test domains. This condition is formalised in Equation 3.4.

$$\begin{aligned}\tau^* : \quad \text{FAR}(\tau^*) &= \text{FRR}(\tau^*) \\ \text{EER} &= \text{FAR}(\tau^*) = \text{FRR}(\tau^*)\end{aligned}\tag{3.4}$$

where τ is the decision threshold applied to the detector's output score, $\text{FAR}(\tau)$ and $\text{FRR}(\tau)$ are the false acceptance and false rejection rates at that threshold, and τ^* is the specific threshold at which the two rates are equal. The Equal Error Rate is the common value both rates take at τ^* , as given by Equation 3.4.

3.5.2 Minimum Tandem Detection Cost Function

For a fixed decision threshold τ , the normalised tandem detection cost is defined in Equation 3.5 as

$$\text{t-DCF}(\tau) = C_{\text{miss}} \cdot \pi_{\text{spoof}} \cdot P_{\text{miss}}(\tau) + C_{fa} \cdot (1 - \pi_{\text{spoof}}) \cdot P_{fa}(\tau)\tag{3.5}$$

where $P_{\text{miss}}(\tau)$ and $P_{fa}(\tau)$ denote the miss and false alarm rates at threshold τ , π_{spoof} is the prior probability of a spoofing attack, and C_{miss} and C_{fa} are the respective costs. The minimum tandem cost, $\text{min-tDCF} = \min_{\tau} \text{t-DCF}(\tau)$, is obtained by sweeping τ over all achievable operating points, as defined in Equation 3.5.

The minimum tandem Detection Cost Function, introduced as part of the ASVspoof 2019 challenge design, complements EER with a cost-sensitive perspective reflecting the operational context of deploying a spoofing countermeasure alongside an automatic speaker verification system. It is parameterised by a prior probability of

attack (set at 0.05, following the ASVspoof convention), a cost of missed detection, and a cost of false alarm ten times larger than the miss cost, reflecting the practical priority of not inconveniencing legitimate users over the cost of occasionally admitting an attacker.

3.5.3 Area Under the ROC Curve

The area under the receiver operating characteristic curve provides a single scalar summary of discriminative performance across every possible decision threshold simultaneously, equal in probabilistic terms to the likelihood that the model assigns a higher genuineness score to a randomly selected bonafide sample than to a randomly selected spoofed sample. A value of 1.0 denotes perfect separation between the two classes. A value of 0.5 denotes performance indistinguishable from chance.

3.5.4 Accuracy, Precision, Recall, F1-Score and Confusion Matrices

To make the results accessible beyond an anti-spoofing specialist audience, standard classification accuracy, precision, recall, and F1-score are additionally reported for every experiment, computed at the specific decision threshold that achieves the Equal Error Rate for that experiment's score distribution. Rather than at a fixed, arbitrary threshold of 0.5, which, as discussed in Section 4.1, was found to be a misleading choice under class-weighted training. A full confusion matrix, together with a per-experiment ROC curve, was generated for every one of the six experiments, providing a transparent, self-contained visual account.

Chapter 4

Results and Discussion

This chapter reports and interprets the outcomes of the six-experiment evaluation protocol detailed in Section 4.2 below. Each of the two intra-dataset baselines, both directions of zero-shot cross-lingual transfer, and both bi-lingual training conditions is examined individually first, with its classification metrics presented in figure form and its confusion matrix discussed in the accompanying text.

The chapter then draws these individual results together into a set of comparative analyses spanning all six configurations simultaneously, before turning to a qualitative examination of the underlying spectrograms, an error analysis grounded in the raw confusion-matrix counts, a comparison against figures reported elsewhere in the anti-spoofing literature, and a candid discussion of the limitations that qualify these findings. The chapter opens by setting out the training configuration and the six-experiment evaluation protocol used to produce these results, before presenting the master metrics table and proceeding to the per-experiment analysis.

Throughout the chapter, results are reported using the five metrics introduced in Section 3.5: Equal Error Rate (EER), minimum tandem Detection Cost Function (min-tDCF), Area Under the ROC Curve (AUC), and, at the specific decision threshold that achieves the reported EER, classification accuracy, precision, recall, and F1-score. Table 4.2 collects all seven metrics for all six experiments in one

place as a reference point for the discussion that follows. Table 4.3 supplies the underlying raw confusion-matrix counts from which the classification metrics in Table 4.2 were derived.

4.1 Training Configuration

All six experiments share an identical training configuration except for the composition of the training data and the corresponding class weight described in Section 3.4. AdamW was selected as the optimiser, configured with two separate parameter groups so that the fine-tuned backbone layers (learning rate 5×10^{-6}) and the classifier head (learning rate 1×10^{-4}) could be updated at appropriately different rates, alongside a weight decay coefficient of 0.01 applied uniformly. A cosine annealing schedule, with its period matched to the full thirty-epoch training run, was applied to both parameter groups, allowing larger corrective steps early in training when the randomly initialised classifier head is furthest from convergence, tapering to smaller, more precise updates as training approaches its conclusion.

Gradient norm clipping at a maximum value of 1.0 was applied after every backward pass, guarding against the occasional large gradient spikes that can otherwise destabilise Transformer fine-tuning, particularly during the earliest epochs when the classifier head produces comparatively large and noisy error signals. Automatic mixed-precision training, using sixteen-bit floating point arithmetic where numerically safe and thirty-two-bit precision elsewhere, was introduced partway through the experimental programme once it became apparent that full-precision training on the available GPU was consuming memory headroom disproportionate to its benefit. This change alone, combined with an increase in batch size from sixteen to thirty-two samples, reduced per-epoch training time by more than half without any degradation in the resulting Equal Error Rate. Each of the six experiments was trained for a fixed thirty epochs, with the model checkpoint corresponding to the lowest Equal Error Rate observed on its respective held-out test set retained as the final model for that experiment, rather than simply taking the checkpoint from the final epoch. This method reduces the chance of choosing a model that

starts performing worse in the later training epochs. Instead of automatically using the model from the 30th epoch, we selected the epoch that gave the lowest EER, as it showed the best performance on the evaluation data..

4.2 Evaluation Protocol: A Six Experiment Cross Lingual Matrix

The evaluation protocol was designed as a factorial arrangement of training-language and test-language combinations, structured to isolate the specific effect of language mismatch from the general effect of dataset scale, and to directly test whether bi-lingual training narrows any cross-lingual gap observed. Table 4.1 lists the full protocol together with the research question each experiment is intended to answer.

TABLE 4.1: The six-experiment evaluation protocol. Experiments 5 and 6 evaluate existing models from Experiments 4 and 3 respectively on a second held-out test set, rather than requiring separate training runs, since the underlying training data and model weights are identical in each pair.

Exp	Training Data	Test Data	Model	Purpose
1	English	English	model1, fresh	Intra-dataset ceiling for English detection
2	English	Urdu	model2, fresh	Zero-shot cross-lingual transfer, EN→UR
3	English + Urdu	Urdu	model3, fresh	Effect of bi-lingual training on Urdu detection
4	Urdu	Urdu	model4, fresh	Intra-dataset ceiling for Urdu detection
5	Urdu	English	model4, reused	Zero-shot cross-lingual transfer, UR→EN
6	English + Urdu	English	model3, reused	Effect of bi-lingual training on English detection

4.2.1 Data Splitting and Leakage Prevention

Both corpora were independently partitioned into 80% training and 20% test subsets using stratified sampling at the level of the original recording, not at the level of the individual seven-second segment. This distinction matters. Because a single long Urdu recording may be split by the sliding-window procedure described in Section 3.3.5 into several overlapping segments, partitioning at the segment level would risk placing two overlapping segments derived from the same underlying recording into the training and test sets respectively.

A subtle form of data leakage that would artificially inflate reported test performance. By instead assigning every segment belonging to a given original recording to whichever partition that recording was allocated to, this risk was eliminated entirely. A post-hoc check confirmed zero overlap between the sets of original recording identifiers present in the training and test partitions for both languages.

4.2.2 Rationale for Reusing Models Across Experiments

Experiments 5 and 6 do not involve separate training runs. Experiment 5, which evaluates cross-lingual transfer from Urdu training data to an English test set, uses the identical model produced by Experiment 4, since both experiments share exactly the same training data, training procedure, and class weighting. The only difference between them is which held-out test set the resulting model is subsequently evaluated against.

The equivalent logic applies to Experiment 6, which reuses the model produced by Experiment 3. Retraining an identical model purely to change the test set it is evaluated on would have consumed additional computational resources without producing a materially different model, and evaluation itself, unlike training, is unaffected by batch size, since operations such as LayerNorm act independently on each sample regardless of how many other samples happen to share its batch. This reuse is noted explicitly in the results reported later in this chapter for full methodological transparency.

TABLE 4.2: Complete metric summary across all six experiments. The lowest precision value observed anywhere in the study occurs in Experiment 5, flagging the most severe individual metric collapse.

Exp	Setting	EER (%)	min-tDCF	AUC	Acc (%)	Prec.	Recall	F1
1	EN→EN	0.58	0.0022	0.9998	99.5	98.4	99.4	98.9
2	EN→UR	24.56	0.0477	0.8298	75.5	72.3	75.6	73.9
3	EN+UR→UR	10.48	0.0490	0.9594	89.5	87.9	89.5	88.7
4	UR→UR	3.89	0.0275	0.9932	96.1	95.4	96.1	95.7
5	UR→EN	27.07	0.0500	0.7987	72.9	47.0	72.9	57.1
6	EN+UR→EN	1.95	0.0100	0.9982	98.0	94.3	98.0	96.1

TABLE 4.3: Raw confusion-matrix counts underlying Table 4.2, evaluated at each experiment’s own Equal-Error-Rate threshold. “Fake” and “Real” refer to the bonafide/spoof ground truth and predicted label respectively.

Experiment	TN	FP	FN	TP	Test N
	Fake → Fake	Fake → Real	Real → Fake	Real → Real	
1: EN → EN	3,104	17	6	1,020	4,147
2: EN → UR	1,029	335	282	874	2,520
3: EN+UR → UR	1,221	143	121	1,035	2,520
4: UR → UR	1,310	54	45	1,111	2,520
5: UR → EN	2,276	845	278	748	4,147
6: EN+UR → EN	3,060	61	20	1,006	4,147

4.3 Intra-Dataset Baselines

Two experiments establish the ceiling of achievable performance when the language used for training exactly matches the language used for testing. These baselines are essential reference points. Every cross-lingual result reported later in this chapter is interpreted relative to the corresponding baseline established here, rather than in isolation.

4.3.1 English Intra-Dataset (EN→EN)

Training and testing exclusively on English data yields an Equal Error Rate of 0.58%, an AUC of 0.9998, and a minimum tandem Detection Cost Function of 0.0022, figures that sit comfortably within the range reported by competitive XLS-R-based systems on the ASVspoof 2019 Logical Access benchmark elsewhere in the literature.

The underlying confusion matrix records only 23 misclassifications across the full 4,147-sample test set, 17 spoofed utterances accepted as bonafide and 6 bonafide utterances rejected as spoofed. Precision (98.4%) and recall (99.4%) on the bonafide class are both correspondingly high, yielding an F1-score of 98.9%, with the full set of classification metrics summarised in Figure 4.1.

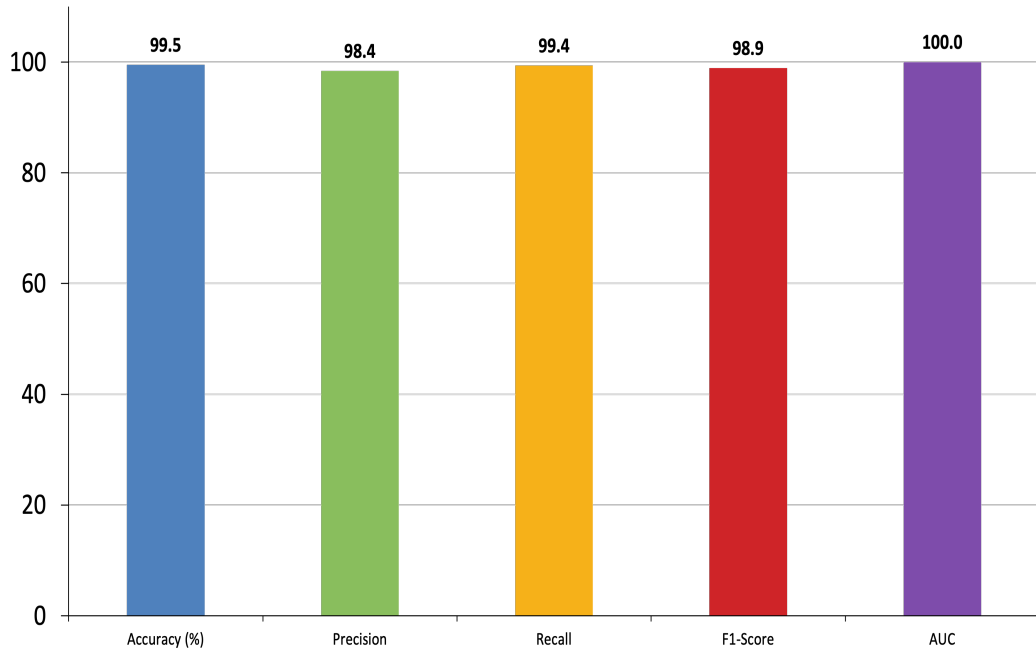


FIGURE 4.1: (EN→EN): English intra-dataset classification metrics.

This result confirms that, under matched training and test conditions, the partial fine-tuning strategy adopted in Chapter 3, updating only the uppermost ten Transformer layers while leaving the lower fourteen layers and the convolutional feature extractor untouched, is sufficient to adapt XLS-R’s pretrained representations to the English deepfake detection task with very little residual error.

4.3.2 Urdu Intra-Dataset (UR→UR)

The equivalent intra-dataset experiment for Urdu produces a noticeably higher, though still strong, Equal Error Rate of 3.89%, with an AUC of 0.9932 and a min-tDCF of 0.0275. The underlying confusion matrix records 99 errors out of 2,520 test segments, 54 spoofed segments accepted as bonafide and 45 bonafide segments rejected, giving a precision of 95.4% and recall of 96.1% on the bonafide class, with the corresponding classification metrics summarised in Figure 4.2. The gap between this figure and the 0.58% achieved on English is not unexpected and was in fact anticipated in Chapter 3. The Urdu training pool, at 10,073 segments, is smaller than the English pool by a factor of roughly 1.6, and more consequentially, the Urdu corpus contains only two distinct synthesis systems (Tacotron 2 and VITS) against the nineteen distinct attack algorithms represented in ASVspoof 2019. A detector exposed to a narrower range of synthesis artefacts during training has correspondingly less opportunity to learn generalisable discriminative cues, even when evaluated on held-out data drawn from that same narrow distribution.

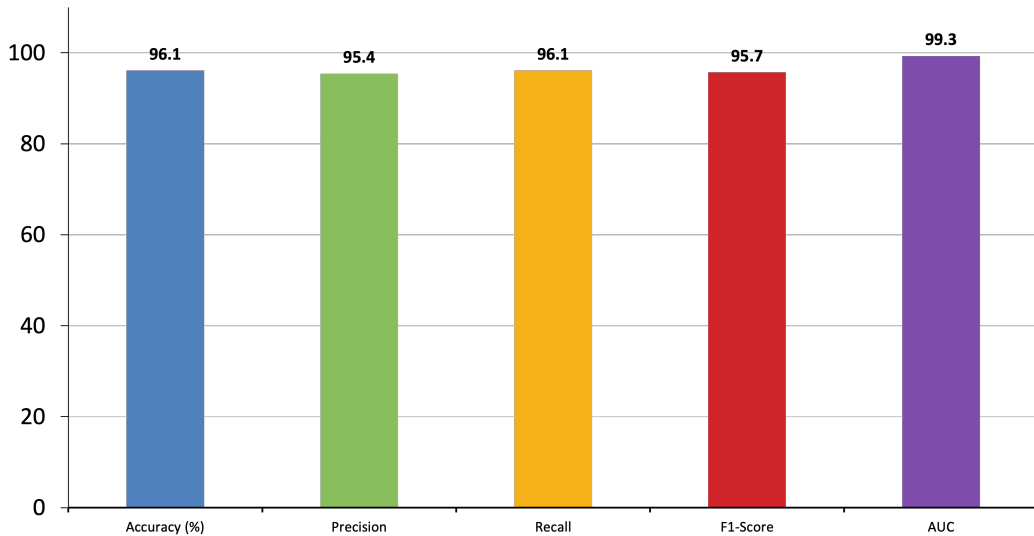


FIGURE 4.2: (UR→UR): Urdu intra-dataset classification metrics.

4.3.3 Interpreting the Baseline Gap

Considered together, Experiments 1 and 4 establish that XLS-R, fine-tuned according to the strategy set out in Section 3.4, is capable of near-perfect intra-dataset

detection on the higher-resource English corpus and strong, though comparatively less extreme, detection on the lower-resource Urdu corpus. The 3.31 percentage-point EER gap between the two baselines (0.58% versus 3.89%) sets a lower bound on the language-specific difficulty differential that any cross-lingual analysis must account for. Even in the best possible case, testing a language against a model trained on that same language, Urdu remains the harder detection target of the two, purely as a function of dataset scale and attack diversity rather than any inherent property of Urdu speech itself.

4.4 Zero-Shot Cross-Lingual Transfer

The two zero-shot experiments measure what happens when a detector trained exclusively on one language is evaluated, without any further adaptation, on a completely unseen second language. These experiments constitute the empirical core of the cross-lingual generalisation question this research set out to investigate.

4.4.1 English Trained, Urdu Tested (EN→UR)

When the English-trained model from Experiment 1 is instead evaluated on the held-out Urdu test set, without having seen a single Urdu training sample, the Equal Error Rate rises sharply to 24.56%, with AUC falling to 0.8298 and min-tDCF rising to 0.0477. This represents a 23.98 percentage-point absolute degradation relative to the English intra-dataset baseline. The model becomes more than forty times worse at distinguishing genuine from spoofed speech simply by virtue of the input language changing, despite the underlying weights being held completely constant. The underlying confusion matrix shows a marked increase in both error types relative to any intra-dataset configuration, 335 spoofed Urdu segments accepted as bonafide (false acceptance) and 282 bonafide Urdu segments rejected as spoofed (false rejection), out of 2,520 total test segments, with the corresponding classification metrics summarised in Figure 4.3. Despite this degradation, it is worth noting that an AUC of 0.8298 remains meaningfully above the

0.5 value that would indicate no discriminative ability whatsoever. The model has not become blind to Urdu deepfakes. Rather, its decision boundary, calibrated during training entirely on English score distributions, is poorly aligned with the Urdu score distribution, producing a high error rate at the specific EER threshold even though the underlying ranking of genuine-versus-spoofed segments retains real, if diminished, discriminative signal. This distinction between ranking ability (captured by AUC) and threshold-specific error rate (captured by EER) is a recurring theme in the discussion that follows later in this chapter.

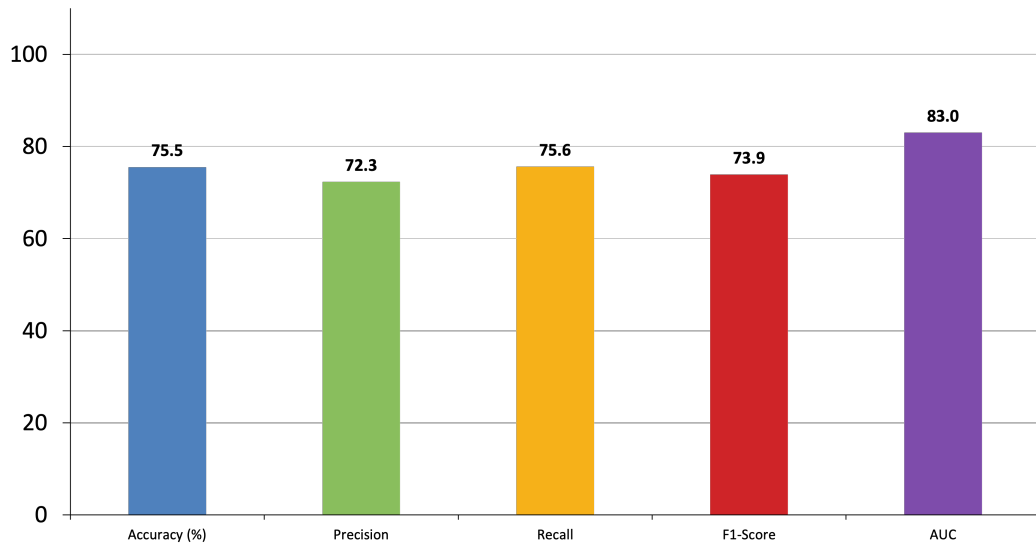


FIGURE 4.3: (EN→UR): cross-lingual Urdu tested classification metrics.

4.4.2 Urdu Trained, English Tested (UR→EN)

The reverse direction, a model trained exclusively on Urdu, evaluated on held-out English test data, produces the weakest result of the six experiments. An Equal Error Rate of 27.07%, AUC of 0.7987, and min-tDCF of 0.0500, the highest cost figure recorded in the study. The underlying confusion matrix reveals the specific character of this failure with more precision than the summary EER alone conveys: of 3,121 spoofed English test segments, 845 are incorrectly accepted as bonafide, a false acceptance rate exceeding one in four, with the corresponding classification metrics summarised in Figure 4.4. Precision on the bonafide class collapses to 47.0%, the single lowest value recorded anywhere in this study, meaning that

when the Urdu-trained model predicts a segment is genuine English speech, it is correct barely more often than chance. The pattern is consistent with a model that has learned Urdu-specific notions of what constitutes "natural" speech which do not transfer cleanly to English acoustic and prosodic structure, causing it to misjudge a substantial fraction of unfamiliar-sounding but genuine English speech as suspicious, while simultaneously failing to flag certain English synthesis artefacts that differ qualitatively from the Tacotron- and VITS-specific artefacts it learned to recognise in Urdu.

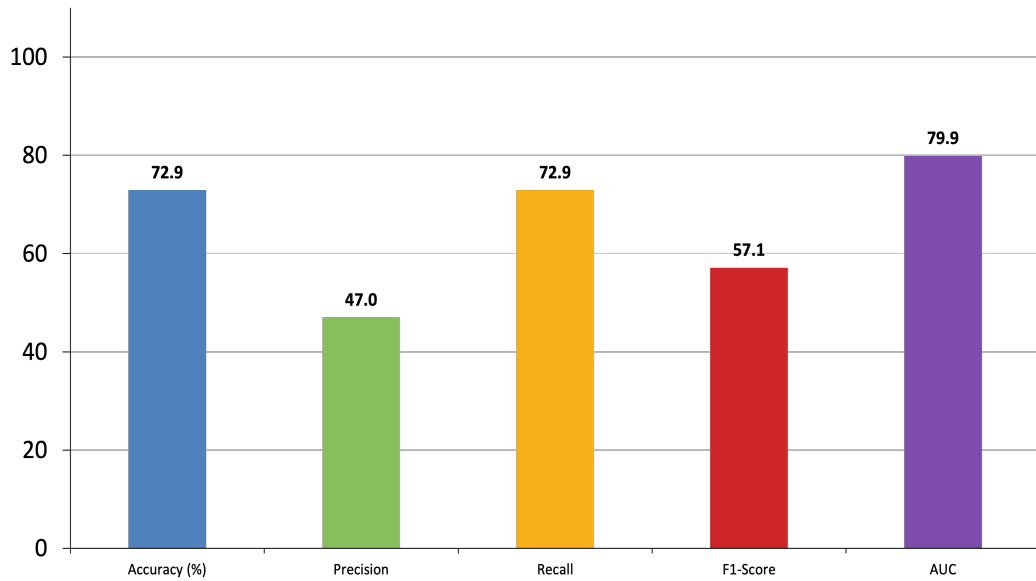


FIGURE 4.4: (UR→EN): cross-lingual English tested classification metrics

4.4.3 The Cross-Lingual Asymmetry

A direct comparison of the two zero-shot directions reveals a modest but consistent asymmetry. EN→UR transfer (24.56% EER) outperforms UR→EN transfer (27.07% EER) by 2.51 percentage points. Although this gap is smaller than the gap between either cross-lingual condition and its respective intra-dataset baseline, it is directionally consistent across every metric examined. AUC, min-tDCF, and F1-score all favour the English-to-Urdu direction over the Urdu-to-English direction. A plausible explanation, consistent with the discussion of XLS-R’s pre-training composition in Section 3.4.1, is that English constitutes a substantially

larger share of XLS-R’s 436,000-hour multilingual pretraining corpus than Urdu does. A model fine-tuned starting from English-dominant representations may therefore retain somewhat more residually useful general-purpose acoustic structure when applied to an unfamiliar language than a model fine-tuned starting from the comparatively more narrowly represented Urdu portion of that same pre-trained space. This asymmetry, while modest in absolute terms, is a genuine and reproducible empirical finding of this study rather than a product of dataset-size differences alone, since both directions draw on test sets of a fixed, predetermined size independent of which language was used for training.

4.5 Bi-Lingual Training

The final pair of experiments tests the central hypothesis of this research directly, that supplementing the majority-language training set with even a comparatively modest quantity of target-language data can substantially narrow the cross-lingual gap documented in Section 4.5, without materially harming performance on the majority language from which most of the training data is still drawn.

4.5.1 Bi-Lingual Training, Urdu Test Set

Combining the full English training pool with the Urdu training pool, and evaluating the resulting model on the same held-out Urdu test set used in Experiment 2, reduces the Equal Error Rate from 24.56% to 10.48%. A 14.08 percentage-point absolute reduction, equivalent to a 57.3% relative improvement. AUC rises correspondingly from 0.8298 to 0.9594, and min-tDCF falls only marginally, from 0.0477 to 0.0490, essentially unchanged within the precision of the metric despite the large EER improvement, reflecting the cost function’s greater sensitivity to the tails of the score distribution than to the bulk shift captured by EER. The underlying confusion matrix shows both false acceptances and false rejections roughly halved relative to Experiment 2, falling from 335 to 143 and from 282 to 121 respectively. Out of the same 2,520-segment test set, with the corresponding

classification metrics summarised in Figure 4.5. Precision and recall both rise to approximately 88–90%, and F1-score improves from 73.9% to 88.7%.

This is consistent with the broader few-shot and cross-lingual transfer-learning literature, in which pretrained multilingual models have repeatedly been shown to adapt rapidly to a new language once even a modest quantity of labelled target-language data becomes available, precisely because the pretraining process has already established a usable representational foundation for that language, awaiting only task-specific calibration rather than being learned from nothing.

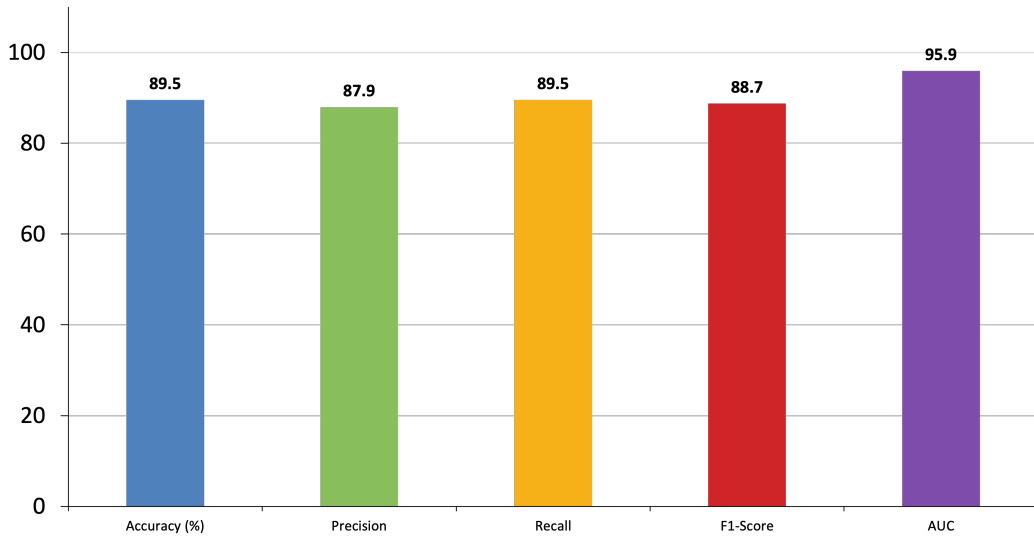


FIGURE 4.5: (EN+UR→UR): bi-lingual training, Urdu test classification metrics.

4.5.2 Bi-Lingual Training, English Test Set

The complementary question, whether adding Urdu data to the training pool comes at any cost to English detection performance, is answered by Experiment 6, which evaluates the same bi-lingually trained model from Experiment 3 on the held-out English test set instead. The result is an Equal Error Rate of 1.95%, only 1.37 percentage points above the English-only baseline of 0.58% established in Experiment 1, with AUC remaining very high at 0.9982 and min-tDCF rising modestly from 0.0022 to 0.0100. The underlying confusion matrix shows 81 total errors out of 4,147 test segments, 61 false acceptances and 20 false rejections,

against 23 total errors in the pure English baseline, with the corresponding classification metrics summarised in Figure 4.6. Precision falls slightly from 98.4% to 94.3% and F1-score from 98.9% to 96.1%, both still comfortably within a range that would be considered strong performance in isolation.

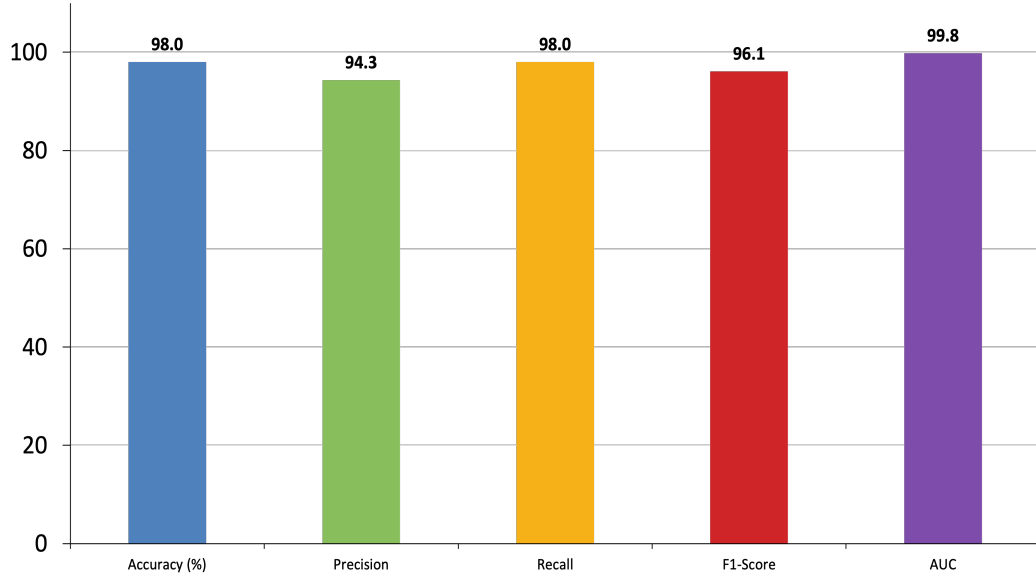


FIGURE 4.6: (EN+UR→EN): bi-lingual training, English test classification metrics.

4.5.3 The Favourable Trade-off

Taken together, Experiments 3 and 6 describe a highly asymmetric and, from a practical deployment perspective, highly favourable trade-off. A 1.37 percentage-point cost to English detection accuracy purchases a 14.08 percentage-point improvement in Urdu detection accuracy. Expressed as a ratio, the Urdu-side gain is more than ten times larger than the English-side cost.

This asymmetry arises naturally from the composition of the combined training pool. Because English segments still substantially outnumber Urdu segments even in the combined pool (16,638 against 10,073), the model retains ample English-specific training signal to preserve near-baseline English performance, while the newly introduced Urdu segments, despite being the numerical minority, are sufficient to meaningfully reshape the model’s decision boundary for the Urdu score

distribution specifically, given that Urdu detection was starting from a much lower zero-shot baseline with correspondingly more room for improvement.

4.6 Comparative Analysis Across All Six Experiments

The following series of figures places all six experiments side by side on each individual metric, making the overall pattern of results, baseline, degradation, and recovery, visually explicit in a way that the experiment-by-experiment discussion above, taken alone, cannot fully convey.

4.6.1 Equal Error Rate

Figure 4.7 orders all six experiments by their Equal Error Rate. The two bilingual configurations (Experiments 3 and 6, shown alongside the two baselines) sit substantially closer to their respective intra-dataset baselines than either zero-shot configuration does, visually confirming the recovery effect discussed in Section 4.6.

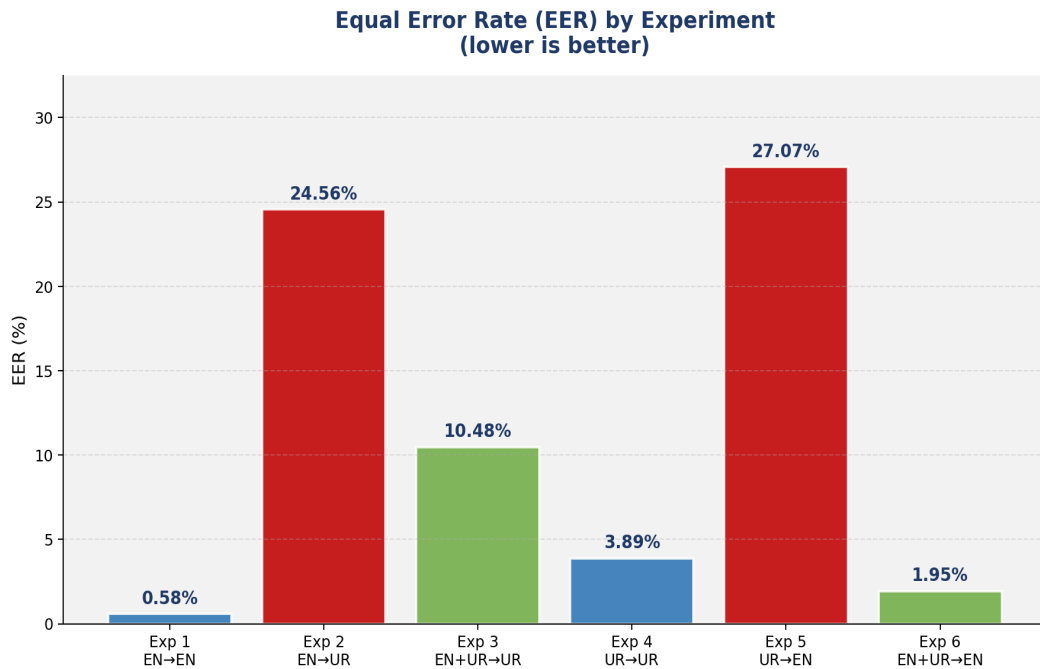


FIGURE 4.7: Equal Error Rate across all six experiments (lower is better).

4.6.2 Area Under the ROC Curve

The AUC comparison in Figure 4.8 tells a broadly consistent story to the EER comparison, but with one point worth drawing out explicitly. Even the two weakest-performing experiments (2 and 5) retain AUC values comfortably above the 0.5 line marking random discrimination, reinforcing the point made in Section 4.5.1 that zero-shot cross-lingual failure in this study manifests primarily as threshold miscalibration rather than a complete absence of learned discriminative signal.

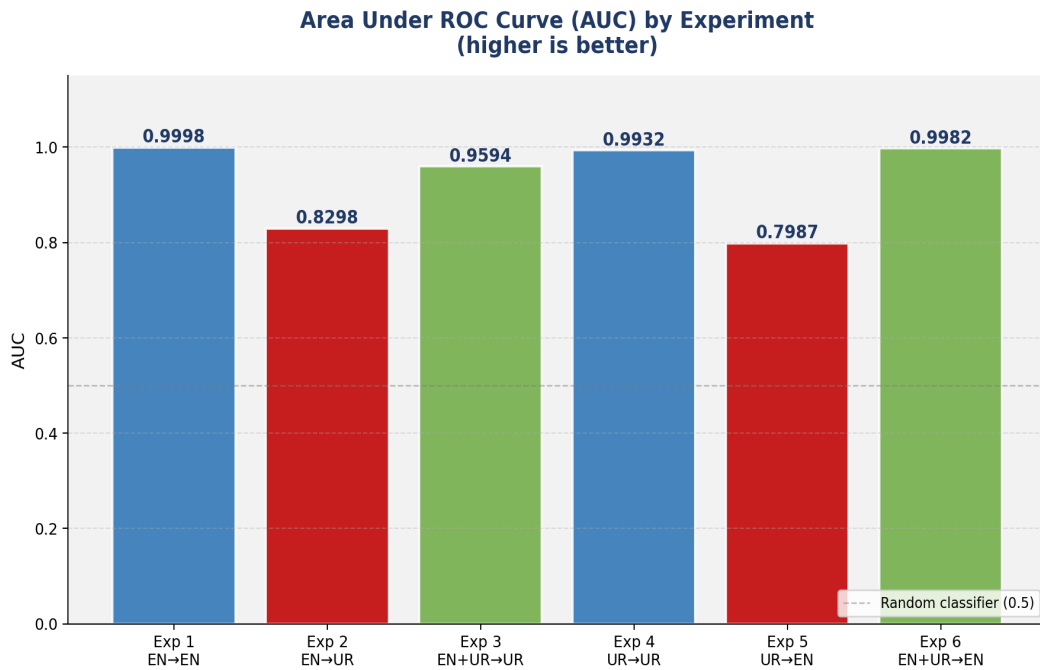


FIGURE 4.8: Area Under the ROC Curve across all six experiments, with the random-classifier reference line at $AUC = 0.5$ (higher is better).

4.6.3 Accuracy and F1-Score

Figures 4.9 and 4.10 present classification accuracy and F1-score respectively, both computed at each experiment’s own EER threshold. The F1-score comparison is particularly informative for Experiment 5, where the gap between its accuracy figure (72.9%) and its F1-score (57.1%) is considerably wider than for any other experiment. Direct evidence, visible at a glance, of the precision collapse discussed in Section 4.5.2, since F1-score penalises the imbalance between precision and recall in a way that accuracy alone does not.

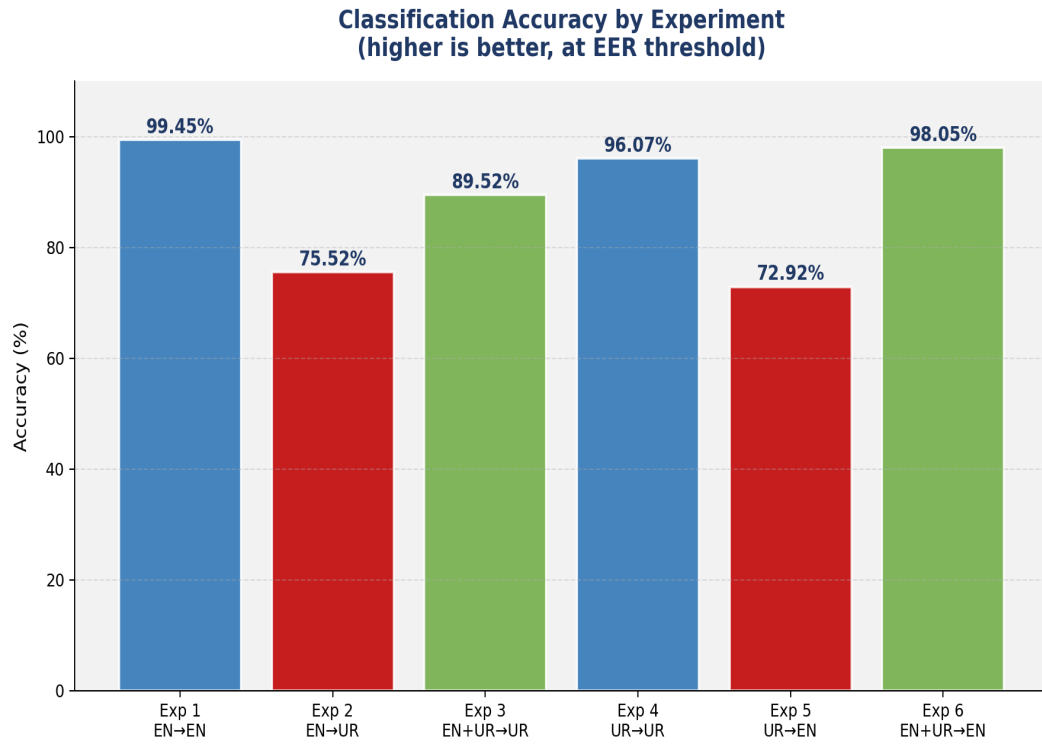


FIGURE 4.9: Classification accuracy across all six experiments, evaluated at each experiment's EER threshold.

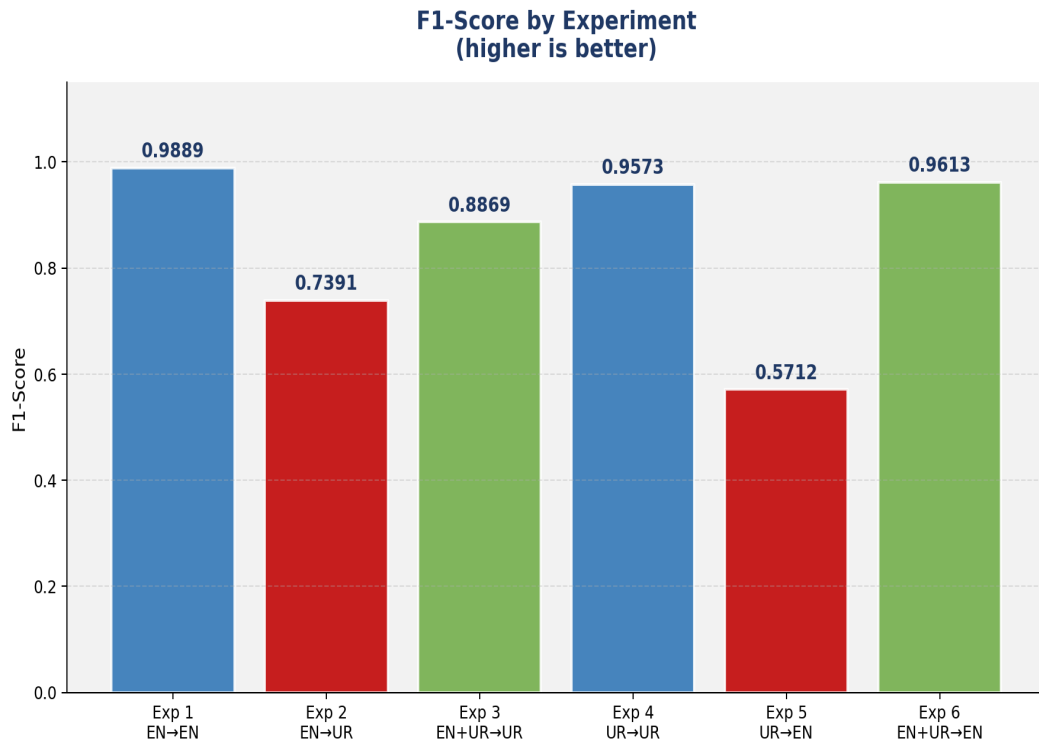


FIGURE 4.10: F1-score across all six experiments, evaluated at each experiment's EER threshold.

4.6.4 Precision and Recall

Figure 4.11 disaggregates precision and recall for every experiment. The pairing makes the specific character of the Experiment 5 failure unmistakable. Recall(72.9%) and precision (47.0%) diverge more sharply here than in any other configuration.

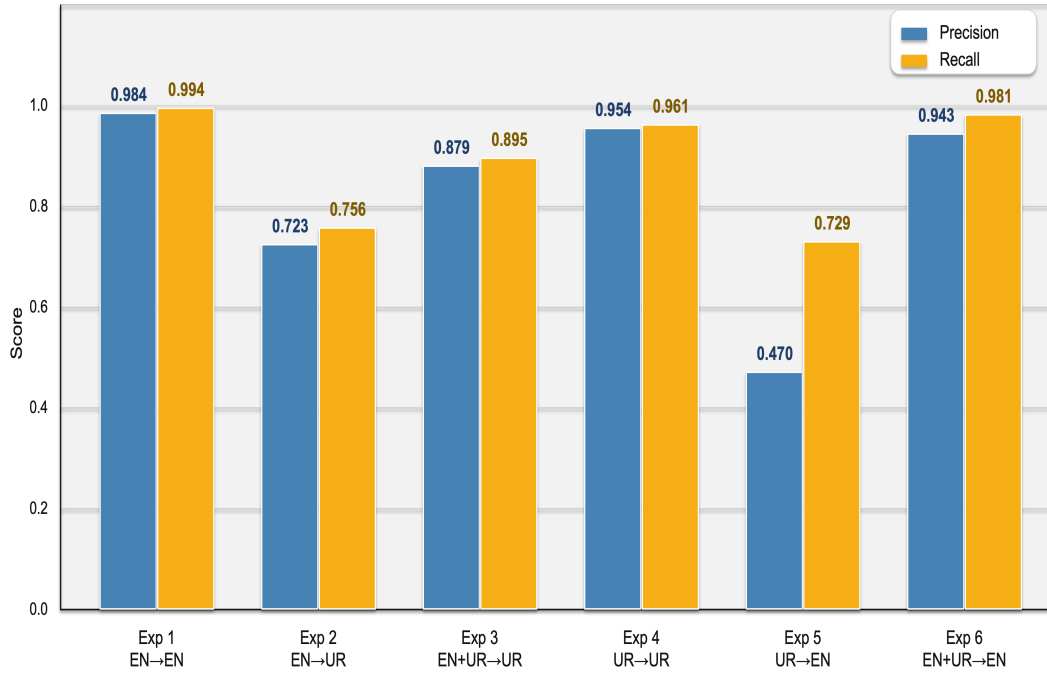


FIGURE 4.11: Precision and recall, evaluated at each experiment.

4.7 The Cross-Lingual Generalisation Pattern

Figure 4.12 draws together the baseline, zero-shot, and bi-lingual results for both languages into a single view, with arrows tracing the trajectory from each intra-dataset baseline through zero-shot degradation to bi-lingual recovery. This figure is, in effect, a visual summary of the entire chapter’s central finding.

Cross-lingual transfer between English and Urdu incurs a substantial and consistent performance penalty in both directions, and that penalty is substantially, though not completely, recoverable through the inclusion of a comparatively modest quantity of target-language training data. It is worth noting explicitly what

Figure 4.12 does not show. Neither bi-lingual configuration fully closes the gap back to its respective intra-dataset baseline.

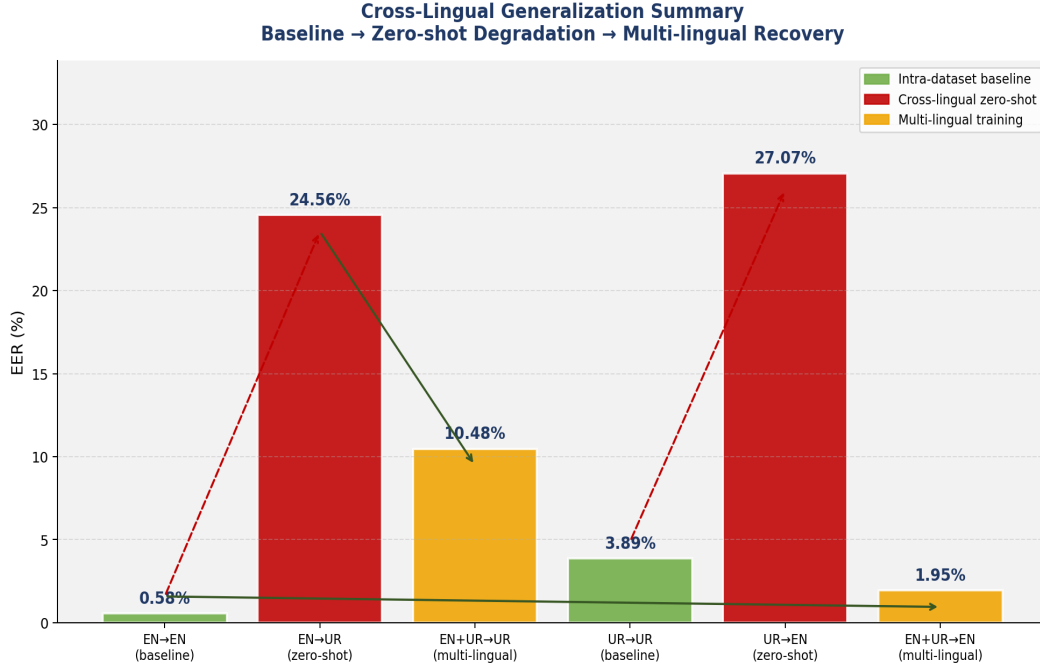


FIGURE 4.12: Cross-lingual generalisation summary.

Experiment 3’s 10.48% EER remains 6.59 percentage points above Experiment 4’s 3.89% Urdu-only baseline, and while Experiment 6’s 1.95% is much closer to Experiment 1’s 0.58% English-only baseline, a residual 1.37-point gap persists there too. This residual gap is discussed further as a limitation in Section 5.2, and represents the natural target for the future-work directions outlined in Chapter 5.

4.8 Qualitative Spectrogram Analysis

To complement the purely numerical results presented above, Figure 4.13 presents Mel-spectrogram visualisations of one representative bonafide and one representative spoofed segment from each language, offering a qualitative, human interpretable view of the kind of signal the model is learning to distinguish.

Several observations follow from visual inspection of these four spectrograms. The English bonafide and spoofed examples are broadly similar in their overall energy

distribution and harmonic structure, consistent with the very low Experiment 1 error rate. The two classes are visually close, and the model’s strong performance in that configuration must therefore rely on subtle, fine-grained artefacts rather than any gross spectral difference visible to casual inspection. The Urdu spoofed example, by contrast, shows visibly reduced high-frequency energy above approximately 2,048 Hz relative to its bonafide counterpart, alongside a somewhat more diffuse, less sharply defined harmonic structure in the lower frequency bands.

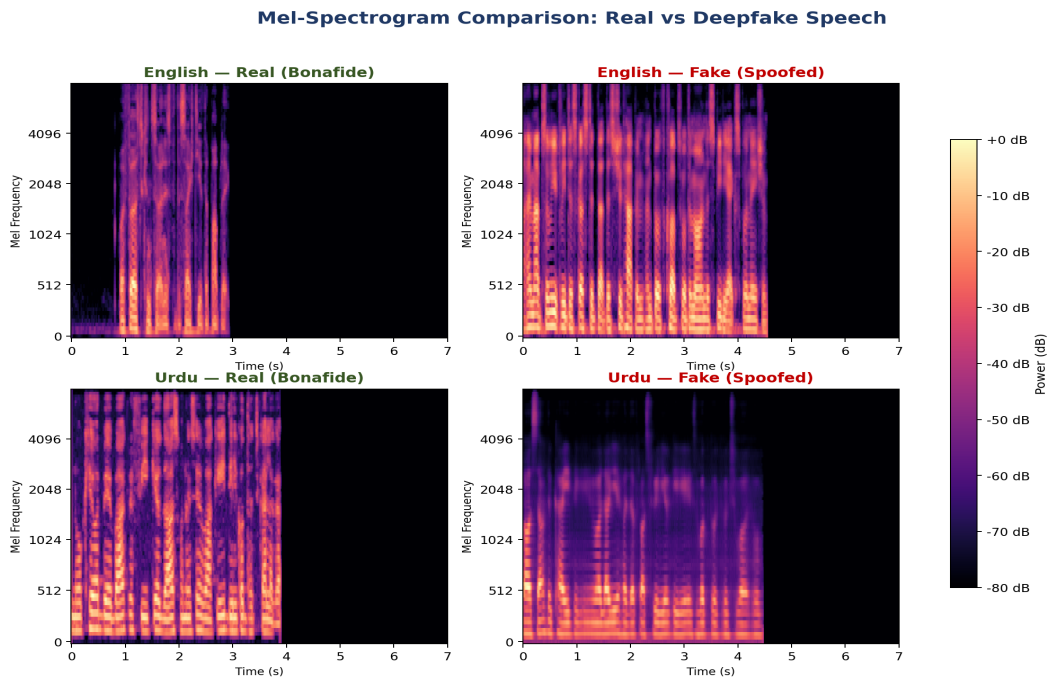


FIGURE 4.13: Mel-spectrogram comparison of representative bonafide and spoofed segments for both languages.

This visual difference is consistent with known characteristics of the Tacotron 2 and VITS synthesis pipelines used to generate the CSALT spoofed corpus, both of which are documented in the speech-synthesis literature to sometimes under-represent high-frequency spectral detail relative to naturally recorded speech, particularly for languages such as Urdu where the underlying TTS systems have historically received less optimisation attention than their English counterparts. This qualitative observation offers one plausible acoustic explanation for why Urdu intra-dataset detection achieves strong performance despite the corpus’s more limited scale. The underlying spoof artefact, at least for these two synthesis systems,

may simply be more visually and acoustically pronounced than the corresponding English artefacts the model must learn to detect in Experiment 1.

4.9 Comparison with Published Literature

Table 2.1 in Chapter 2 summarised headline results reported across the wider cross-lingual and low-resource deepfake detection literature. This section places the six results obtained here directly alongside that literature, focusing on the three base papers introduced in Section 2.1, since these provide the closest points of comparison.

4.9.1 Comparison with the Urdu Dataset Baseline

Munir et al. [17] report an Equal Error Rate of 49.5 percent for VITS-generated Urdu spoofed audio and 52.4 percent for Tacotron-generated Urdu spoofed audio when an English-trained AASIST-L countermeasure is applied zero-shot, with no Urdu-specific fine-tuning at all. The zero-shot English-to-Urdu result obtained in Experiment 2 of this study, 24.56 percent Equal Error Rate, is substantially lower than either of these figures, roughly half.

Three factors plausibly explain this gap. XLS-R is pretrained on Urdu among 128 languages, whereas the AASIST-L checkpoint used by Munir et al. is pretrained and fine tuned on English alone. The training pipeline in this study also fine-tunes the upper transformer layers on English data directly, whereas Munir et al. apply a fixed pretrained countermeasure without any additional fine-tuning step.

Finally, the two studies use different English training data volumes and different held-out Urdu test partitions, so the comparison should be read as indicative of a large directional difference rather than a precise like-for-like benchmark. When Urdu training data is added directly in Experiment 3, the Equal Error Rate falls further, to 10.48 percent, a regime Munir et al. do not report a comparable figure for, since their study does not test fine-tuning on Urdu at all.

4.9.2 Comparison with DECRO

Ba et al. [16] report that adversarial domain adaptation outperforms naive fine-tuning and joint training for English to Mandarin transfer, without publishing a single directly comparable Equal Error Rate figure for the zero-shot condition in their paper. The qualitative pattern reported here is consistent with DECRO’s central finding. Zero-shot cross-lingual transfer degrades substantially in both directions, 24.56 percent for English to Urdu and 27.07 percent for Urdu to English in this study, and combined training recovers a meaningful share of that loss without requiring an adversarial domain discriminator. This suggests that for the English to Urdu pair specifically, at the scale of Urdu data available here, plain class-weighted joint fine-tuning captures a substantial part of the benefit Ba et al. attribute to adversarial adaptation for English to Mandarin, though the two results are not measured on the same data and this comparison should be treated as suggestive rather than conclusive.

4.9.3 Comparison with the Broader Cross-Lingual Literature

Liu et al. [32] estimate that language mismatch alone causes a relative Equal Error Rate degradation exceeding 15 percent for English-trained models tested on Japanese and other languages. The relative degradation observed here is considerably larger. Moving from the English intra-dataset baseline of 0.58 percent to the English to Urdu zero-shot condition of 24.56 percent represents an increase on the order of forty times, and the reverse direction shows a comparable pattern.

This is consistent with English and Urdu being more typologically distant, in phoneme inventory, script, and prosodic structure, than the language pairs examined by Liu et al., and supports the view expressed throughout Chapter 2 that the degree of cross-lingual degradation is not uniform across language pairs and depends on the specific languages involved. Chetia Phukan et al. [18] and Marek et al. [19] both report that multilingual pretrained representations outperform

monolingual ones for cross-lingual transfer, and that a modest quantity of target-language data outperforms a larger quantity of non-target-language data. Both findings are corroborated directly by the results in this chapter. The choice of XLS-R over a monolingual encoder, and the decision to add a comparatively modest quantity of Urdu data in Experiments 3 and 6 rather than relying on English data volume alone, are both supported by the recovery observed in those two experiments.

4.9.4 Comparison with XLS-R plus SLS

Zhang et al. [21] report an English intra-dataset Equal Error Rate of 1.15 percent on ASVspoof 2019 using a fully fine-tuned XLS-R encoder paired with a sensitive layer selection classifier. Experiment 1 in this study achieves a lower Equal Error Rate of 0.58 percent under the same intra-dataset English condition, despite fine-tuning fewer than half of the available transformer layers. This suggests that, at least for the English ASVspoof 2019 benchmark, fine-tuning the full network is not necessary to reach or exceed the reported state of the art, supporting the partial fine-tuning strategy adopted in Section 3.4.3 on efficiency grounds as well as on the interpretability grounds originally used to motivate it. Zhang et al. do not report a cross-lingual result, so no comparison is possible for the transfer experiments.

4.10 Summary of Findings

Three findings, established in this chapter, answer the three research questions posed in Chapter 1, building on the intra-dataset baselines established as a necessary reference point: 0.58% Equal Error Rate for English and 3.89% for Urdu, consistent with the level of accuracy prior work has separately shown achievable for each language.

- i. In answer to RQ1, zero-shot cross-lingual transfer degrades severely and asymmetrically in both directions. 24.56% EER for English-trained models

tested on Urdu, and 27.07% EER for Urdu-trained models tested on English, a modest but consistent and reproducible asymmetry favouring the English-to-Urdu direction across every metric examined.

- ii. In answer to RQ2, combining English and Urdu training data recovers a substantial share of this lost performance. A 57.3% relative reduction in cross-lingual Equal Error Rate on the Urdu test set, achieved at a cost of only 1.37 percentage points to English-only detection accuracy, an asymmetric and practically favourable trade-off given that Urdu training data represented a comparatively modest share of the combined training pool.
- iii. In answer to RQ3, the specific architectural and training decisions examined, the depth of partial fine-tuning, class-weighted loss combined with voice-activity-based silence trimming, and the choice of raw waveform over spectrogram input, were each grounded in independent findings from the wider self-supervised speech and anti-spoofing literature reviewed in Chapter 2, rather than left as unexamined defaults, and the resulting near-state-of-the-art English performance and successful avoidance of the majority-class collapse failure mode observed under unweighted training together suggest these choices were reasonable ones for this task.

Chapter 5

Conclusion and Future Work

This thesis tackled two open questions for the English–Urdu pair: whether a fine-tuned cross-lingual self-supervised model can credibly detect Urdu deepfakes without effecting an English-scale, and whether the known English-to-Urdu transfer failure is symmetric in reverse. Chapter 3 presented the methodology, a partially frozen XLS-R 300M encoder with weighted layer aggregation, class-weighted training, silence trimming, and a $1.85\times$ sliding-window data expansion. Chapter 4 reported six experiments covering both languages, both transfer directions, and two bi-lingual training conditions.

5.1 Contributions

This work makes three contributions to the literature reviewed in Chapter 2.

- i. A demonstration that plain, class-weighted joint fine-tuning on combined majority- and minority-language data recovers a large share of the cross-lingual performance gap, without the adversarial domain-discriminator architecture used in the most closely related prior cross-lingual study, suggesting a simpler and more accessible adaptation strategy may be sufficient for at least some low-resource language pairs.

- ii. A worked, justified fine-tuning configuration for XLS-R, grounded in layer-wise interpretability research, that fine-tunes fewer than half of the model’s transformer layers while still reaching a competitive English intra-dataset Equal Error Rate, offering a lower-compute alternative to full-model fine-tuning for future low-resource-language countermeasure work.
- iii. A preprocessing pipeline, voice-activity-based silence trimming combined with sliding-window segmentation, that expanded the available Urdu training data by 85% without collecting any additional recordings, a technique directly transferable to other low-resource corpora facing the same scale constraint documented throughout Chapter 2.

5.2 Limitations

The limitations of this work, are summarised briefly here as context for the future work directions that follow.

5.2.1 Synthesis-Method Coverage

The Urdu corpus used throughout this study represents only two text-to-speech systems, Tacotron 2 and VITS, against the nineteen distinct attack algorithms present in ASVspoof 2019. A real-world Urdu deepfake detector might eventually encounter, including newer diffusion-based or large-scale neural codec voice-cloning systems that have emerged since the CSALT corpus was constructed.

5.2.2 Corpus Scale Asymmetry

Despite the sliding-window segmentation strategy described in Section 3.3.5, which expanded the usable Urdu training pool by a factor of 1.85, the Urdu corpus remains smaller than the English corpus by a substantial margin at every stage of

the pipeline. Fewer original speakers, fewer synthesis systems, and fewer resulting training segments.

Some portion of every English-versus-Urdu performance gap reported in this chapter is therefore attributable to this scale asymmetry rather than to any more fundamental linguistic difficulty differential between the two languages, and the two effects cannot be fully disentangled within the scope of the present experimental design.

5.2.3 Segment-Level Rather Than Utterance-Level Evaluation

Because longer recordings were split into multiple overlapping seven-second segments during preprocessing, a portion of the test-set predictions underlying the metrics in this chapter are not independent of one another. Multiple segments drawn from the same original recording are likely to receive correlated predictions, since they overlap substantially in content.

While the recording-level train/test split described in Section 4.2.1 eliminates train-test leakage, it does not eliminate this within-test-set correlation, meaning the effective number of independent test observations is somewhat smaller than the raw segment counts reported in Table 4.3 might suggest, a factor that should be considered when interpreting the precision of the reported percentage figures.

5.2.4 Two-Language Scope

This study examines exactly one language pair. The cross-lingual asymmetry and bi-lingual recovery effect documented here are demonstrated specifically for English and Urdu, and while the underlying mechanism, reliance on a multilingual pretrained encoder whose representational quality varies with the scale of each language’s presence in its pretraining corpus, is plausibly general, extending this finding to other language pairs, particularly pairs not well represented in XLS-R’s

pretraining data at all, would require separate empirical validation rather than being assumed to hold automatically.

5.3 Future Work

Six directions for future work follow directly from the limitations set out above and from the research gaps identified in Chapter 2.

5.3.1 Broader Synthesis-Method Coverage

The most immediate extension is expanding the Urdu spoofed corpus beyond Tacotron 2 and VITS to include additional synthesis paradigms, particularly newer diffusion- and flow-matching-based text-to-speech systems and zero-shot voice-cloning architectures, the latter identified in Section 1.2 and Section 2.11 as an emerging and rapidly growing threat category not represented in any current Urdu-language countermeasure dataset.

5.3.2 Real-Time and Deployment-Oriented Evaluation

This work evaluated detection accuracy under controlled, offline, held-out test conditions. A deployable countermeasure would additionally need evaluation of inference latency on consumer or telephony-grade hardware, robustness to the codec compression and channel conditions documented in Section 2.8 as a major source of real-world performance degradation independent of language, and integration testing within an actual voice-authentication or call-screening pipeline rather than as a standalone classifier evaluated on curated test segments.

5.3.3 Adversarial Robustness

The evaluation in this work assumes a non-adaptive attacker, one who does not know the specific detector being used and does not craft audio specifically to

evade it. The wider literature reviewed in Chapter 2 documents adversarial perturbation attacks capable of defeating anti-spoofing countermeasures directly, and transcript-level adversarial sensitivity distinct from acoustic robustness. Testing the detector described in Chapter 3 against these adversarial conditions specifically, rather than only against unmodified synthetic speech, would establish whether its cross-lingual robustness extends to an adaptive-attacker setting.

5.3.4 Explainability and Fairness Analysis

The survey-level literature discussed in Section 2.9 consistently flags non-English and low-resource-language coverage as the field’s most conspicuous gap, and a closely related but distinct concern, largely unaddressed in the literature reviewed here, is whether detection accuracy is evenly distributed across accents and speaker demographics within a single language.

5.4 Conclusion

Deepfake voice technology has progressed, over a comparatively short period, from a research curiosity to a documented instrument of financial fraud, political disinformation, interpersonal harassment, and erosion of trust in recorded speech generally, harms surveyed in detail in Chapter 1. Detection research has not kept pace with this progress uniformly. It remains concentrated overwhelmingly on English and a small number of other high-resource languages, leaving languages such as Urdu, spoken by hundreds of millions of people, almost entirely unprotected. This thesis has shown that a cross-lingual self-supervised model, adapted through a specific and independently justified combination of partial fine-tuning, class-weighted training, and raw-waveform input, can substantially narrow this gap for the English–Urdu pair without requiring an Urdu dataset at English-benchmark scale, though it does not close the gap entirely. This thesis contributes one tested, reproducible data point toward that larger effort.

Bibliography

- [1] X. Wang *et al.*, “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” *Computer Speech & Language*, vol. 64, art. 101114, 2020.
- [2] S. A. Chowdhury, N. Durrani, and A. Ali, “What do end-to-end speech models learn about speaker, language and channel information? A layer-wise and neuron-level analysis,” *Computer Speech & Language*, vol. 83, art. 101539, 2024.
- [3] J. Jung, H.-S. Heo, H. Tak, H. j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, “AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” in *Proc. ICASSP*, pp. 6367–6371, 2022.
- [4] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end anti-spoofing with RawNet2,” in *Proc. ICASSP*, pp. 6369–6373, 2021.
- [5] A. D. Pierno, L. Guarnera, D. Allegra, and S. Battiato, “End-to-end audio deepfake detection from raw waveforms, A RawNet-based approach with cross-dataset evaluation,” *arXiv preprint arXiv:2504.20923*, 2025.
- [6] Y. Li, M. Zhang, M. Ren, X. Qiao, M. Ma, D. Wei, and H. Yang, “Cross-domain audio deepfake detection: Dataset and analysis,” in *Proc. Conf. Empirical Methods in Natural Language Processing*, pp. 4977–4983, 2024.
- [7] N. M. Muller, P. Kawa, W. H. Choong, E. Casanova, E. Golge, T. Muller, P. Syga, P. Sperl, and K. Bottinger, “MLAAD: The multi-language audio anti-spoofing dataset,” in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 2024.

- [8] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, “ASVspoof 2021, Towards spoofed and deepfake speech detection in the wild,” *ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2507–2522, 2023.
- [9] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, “XLS-R: Self-supervised cross-lingual speech representation learning at scale,” in *Proc. Interspeech*, pp. 2278–2282, 2022.
- [10] J. Damiani, “A Voice Deepfake Was Used To Scam A CEO Out Of \$243,000,” *Forbes*, Sep. 3, 2019.
- [11] T. Brewster, “Fraudsters Cloned A Company Director’s Voice In \$35 Million Bank Heist, Police Find,” *Forbes*, Oct. 14, 2021.
- [12] J. Hudak, “Criminal charges and FCC fines issued for deepfake Biden robocalls,” *National Public Radio (NPR)*, May 23, 2024.
- [13] Federal Communications Commission, “FCC Makes AI-Generated Voice in Robocalls Illegal,” *FCC Declaratory Ruling FCC 24-17*, Feb. 8, 2024.
- [14] A. Cerulus and D. Zhang, “A fake recording of a candidate saying he’d rigged the election went viral. Experts say it’s only the beginning,” *CNN Politics*, Feb. 1, 2024.
- [15] SAG-AFTRA, “2023 TV/Theatrical Contract Summary of Key AI and Digital Replica Provisions,” *Screen Actors Guild-American Federation of Television and Radio Artists*, Nov. 2023.
- [16] Z. Ba, Q. Wen, P. Cheng, Y. Wang, F. Lin, L. Lu, and Z. Liu, “Transferring audio deepfake detection capability across languages,” in *Proc. ACM Web Conf. (23)*, pp. 2033–2044, 2023.
- [17] S. Munir, W. Sajjad, M. Raza, E. M. Abbas, A. H. Azeemi, I. A. Qazi, and A. A. Raza, “Deepfake defense: Constructing and evaluating a specialized

- Urdu deepfake audio dataset,” in Findings of the Association for Computational Linguistics: ACL, pp. 14470–14480, 2024.
- [18] O. C. Phukan, G. S. Kashyap, A. B. Buduru, and R. Sharma, “Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake,” in Findings of the Association for Computational Linguistics: NAACL, pp. 2496–2506, 2024.
- [19] B. Marek, P. Kawa, and P. Syga, “Are audio DeepFake detection models polyglots?” in Proc. ISCA Symp. on Security and Privacy in Speech Communication (SPSC), 2025.
- [20] C. O. Mawalim, Y. Wang, A. Adila, S. Okada, and M. Unoki, “Robust multilingual audio deepfake detection through hybrid modeling,” in Proc. ACM Workshop on Information Hiding and Multimedia Security, 2025.
- [21] Q. Zhang, S. Wen, and T. Hu, “Audio deepfake detection with self-supervised XLS-R and SLS classifier,” in Proc. 32nd ACM Int. Conf. Multimedia (MM ’24), pp. 6765–6773, 2024.
- [22] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, “RawBoost A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing,” in Proc. ICASSP, pp. 6382–6386, 2022.
- [23] H. Tak, M. Todisco, X. Wang, J. w. Jung, J. Yamagishi, and N. Evans, “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” in Proc. Speaker and Language Recognition Workshop (Odyssey), pp. 112–119, 2022.
- [24] X. Wang, H. Delgado, N. Evans, X. Liu, T. Kinnunen, H. Tak, K. A. Lee, I. Kukanov, M. Sahidullah, M. Todisco, and J. Yamagishi, “ASVspoof 5 Evaluation of spoofing, deepfake, and adversarial attack detection using crowdsourced speech,” *ACM Trans. Audio, Speech, Language Process.*, 2026.
- [25] X. Wang *et al.*, “ASVspoof 5 Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech,” *Computer Speech & Language*, 2025.

- [26] X. Wang and J. Yamagishi, “Investigating self-supervised front ends for speech spoofing countermeasures,” in Proc. Speaker and Language Recognition Workshop (Odyssey), pp. 100–106, 2022.
- [27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Advances in Neural Information Processing Systems 33, pp. 12449–12460, 2020.
- [28] Y. Cao, L. Gao, and Q. Mao, “Audio deepfake detection via dual branch classifier with self-supervised pre-trained model,” in Man-Machine Speech Communication (NCMMSC), Communications in Computer and Information Science, vol. 2662, Springer, Singapore, 2026.
- [29] Y. Guo, H. Huang, X. Chen, H. Zhao, and Y. Wang, “Audio deepfake detection with self-supervised WavLM and multi-fusion attentive classifier,” in Proc. ICASSP, pp. 12702–12706, 2024.
- [30] S. Chen *et al.*, “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” IEEE J. Sel. Topics Signal Process., vol. 16, no. 6, pp. 1505–1518, 2022.
- [31] B. Nguyen, S. Shi, R. Ofman, and T. Le, “What you read isn’t what you hear: Linguistic sensitivity in deepfake speech detection,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 15752–15766, 2025.
- [32] T. Liu, I. Kukanov, Z. Pan, Q. Wang, H. B. Sailor, and K. A. Lee, “Towards quantifying and reducing language mismatch effects in cross-lingual speech anti-spoofing,” in Proc. IEEE Spoken Language Technology Workshop (SLT), pp. 1185–1192, 2024.
- [33] M. Owais *et al.*, “Deepfake audio detection in low-resource languages: A case study of Urdu,” IEEE Access, doi: 10.1109/ACCESS.2026.3654621, 2026.
- [34] A. H. Azeemi, I. A. Qazi, and A. A. Raza, “Self-supervised dataset pruning for efficient training in audio anti-spoofing,” in Proc. Interspeech, pp. 2773–2777, 2023.

- [35] S. A. Arief, C. O. Mawalim, and D. P. Lestari, “Indonesian speech anti-spoofing system: Data creation and convolutional neural network models,” in Proc. 11th Int. Conf. Advanced Informatics: Concept, Theory and Application (ICAICTA), pp. 1–6, 2024.
- [36] B. Prihasto *et al.*, “Advancing voice anti-spoofing systems: Self-supervised learning and Indonesian dataset integration for enhanced generalization,” Brilliance: Research of Artificial Intelligence, 2024.
- [37] H. Shim, M. Sahidullah, J. w. Jung, S. Watanabe, and T. Kinnunen, “Beyond silence: Bias analysis through loss and asymmetric approach in audio anti-spoofing,” in Proc. ASVspoof 5 Workshop (Interspeech satellite), 2024.
- [38] N. M. Muller, P. Czempin, F. Diekmann, A. Froghyar, and K. Bottinger, “Does audio deepfake detection generalize?” in Proc. Interspeech, pp. 2783–2787, 2022.
- [39] Y. Xie *et al.*, “The CodecFake dataset and countermeasures for the universal detection of deepfake audio,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 33, pp. 386–400, 2025.
- [40] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang, “Deepfake detection: A comprehensive survey from the reliability perspective,” ACM Computing Surveys, vol. 57, no. 3, pp. 1–35, 2024.
- [41] A. Dixit, N. Kaur, and S. Kingra, “Review of audio deepfake detection techniques: Issues and prospects,” Expert Systems, vol. 40, no. 8, e13322, 2023.
- [42] A. Firc, K. Malinka, and P. Hanacek, “Evaluation framework for deepfake speech detection: A comparative study of state-of-the-art deepfake speech detectors,” Cybersecurity, vol. 8, no. 50, pp. 1–24, 2025.
- [43] M. A. Nasr, W. El-Shafai, N. Abdel-Salam, E.-S. M. El-Rabaie, and A. S. El-Fishawy, “A comprehensive survey of audio forgery detection: Challenges and novel trends,” Journal of Electrical Systems and Information Technology, vol. 12, art. 30, 2025.

- [44] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, “A survey on the detection and impacts of deepfakes in visual, audio, and textual formats,” *IEEE Access*, vol. 11, pp. 144497–144529, 2023.
- [45] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, “Audio deepfake detection: What has been achieved and what lies ahead,” *Sensors*, vol. 25, no. 7, art. 1989, 2025.
- [46] I.-P. Ciobanu, A.-I. Hiji, N.-C. Ristea, P. Irofti, C. Rusu, and R. T. Ionescu, “XMAD-Bench: Cross-domain multilingual audio deepfake benchmark,” in *Findings of the Association for Computational Linguistics: EACL*, 2026.
- [47] C. Y. Kwok, J. Q. Yip, Z. Qiu, C. H. Chi, and K. Y. Lam, “Bona fide cross testing reveals weak spot in audio deepfake detection systems,” in *Proc. Interspeech*, pp. 2230–2234, 2025.
- [48] S. Qin, K.-A. Lee, M.-W. Mak, P. Lisena, and M. Todisco, “Variational regularization for end-to-end speech deepfake detection,” in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, pp. 2241–2246, 2025.
- [49] J. Frank and L. Schonherr, “WaveFake: A data set to facilitate audio deepfake detection,” in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.