



## Deepfake audio detection in low-resource languages: A case study of Urdu

Title	Deepfake audio detection in low-resource languages: A case study of Urdu
Author(s)	Owais, Muhammad;Jadoon, Khurram Khan;Sandhu, Ali Imran;Ali, Zaiwar;Mahmood, Zahid;Yahya, Muhammad;Wahid, Abdul
Publication Date	2026-01-16
Publisher	Institute of Electrical and Electronics Engineers
Repository DOI	<a href="https://doi.org/10.1109/ACCESS.2026.3654621">https://doi.org/10.1109/ACCESS.2026.3654621</a>

## RESEARCH ARTICLE

# Deepfake Audio Detection in Low-Resource Languages: A Case Study of Urdu

MUHAMMAD OWAIS<sup>1</sup>, KHURRAM KHAN JADOON<sup>1</sup>,  
ALI IMRAN SANDHU<sup>1</sup>, (Senior Member, IEEE), ZAIWAR ALI<sup>1</sup>,  
ZAHID MAHMOOD<sup>2</sup>, MUHAMMAD YAHYA<sup>3</sup>, AND ABDUL WAHID<sup>4</sup>

<sup>1</sup>Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi 23640, Pakistan

<sup>2</sup>Department of Computer Engineering, COMSATS University Islamabad, Islamabad 45550, Pakistan

<sup>3</sup>Valeo Vision Systems, Galway, H54 Y276 Ireland

<sup>4</sup>Data Science Institute, University of Galway, Galway, H91 AEX4 Ireland

Corresponding author: Abdul Wahid (abdul.a.wahid@universityofgalway.ie)

This work was supported by the Science Foundation Ireland (SFI) under Grant SFI/16/RC/3918 and Grant SFI/12/RC/2289\_P2.

**ABSTRACT** The rapid advancement of Generative Artificial Intelligence (AI) has enabled the creation of highly realistic synthetic audio, presenting significant challenges to digital security, media forensics, and public confidence. While deepfake detection has been extensively explored for high-resource languages like English, low-resource languages remain critically underexamined. This paper introduces a systematic benchmark of deepfake audio detection methods for Urdu, a language characterized by its rich morphology and phonetic complexity. We evaluate convolutional and transformer-based architectures, including LCNN, CNN-LSTM with Attention, and Whisper variants, employing Mel-Frequency Cepstral Coefficients and Linear Frequency Cepstral Coefficients as front-end features. To address data scarcity, we construct three dataset configurations, baseline, augmented, and extended derived from an existing Urdu deepfake corpus and enhanced through controlled augmentation and additional recordings. Our experiments, conducted with multiple random seeds and statistical validation, demonstrate that MFCC-based models, particularly Whisper-small, achieve strong performance with an Equal Error Rate as low as 0.50%. Robustness tests under noise, pitch, and tempo perturbations highlight the limitations of lightweight CNNs and underscore the advantages of transformer embeddings for handling Urdu's linguistic variability. This study represents the first structured benchmark of deepfake audio detection techniques for Urdu, offering empirical insights into how language characteristics influence model performance and generalization. The findings emphasize the importance of multilingual evaluation in the development of trustworthy speech forensics systems.

**INDEX TERMS** Deepfake audio, synthetic speech detection, MFCC, LFCC, CNN-LSTM, LCNN, whisper transformer, equal error rate (EER).

## I. INTRODUCTION

The rapid development of generative artificial intelligence (AI) has enabled the creation of highly realistic synthetic media, including images, videos, and speech. Among these, deepfake audio produced using techniques such as voice cloning, text-to-speech (TTS) synthesis, and speech-to-speech conversion has become a growing security and forensic concern [1], [2], [3], [4]. Synthetic voices can

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato<sup>1</sup>.

convincingly imitate real speakers, creating vulnerabilities in financial fraud, identity verification, political misinformation, and social engineering attacks [5], [6]. Recent advances in deep neural architectures, including autoregressive models such as WaveNet [7] and Generative Adversarial Networks (GANs) [8], have significantly improved the rhythmic and intonational aspects of speech, naturalness, and linguistic fidelity of synthetic speech, complicating reliable detection [9].

Existing deepfake detection research has concentrated primarily on high-resource languages, especially English [10],

[11], [12], [13], [14]. Techniques have ranged from traditional feature-based classifiers, leveraging Mel-Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), or Constant Q Cepstral Coefficients (CQCC) [15], [16], [17], to modern deep learning approaches based on Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models [18], [19], [20], [21], [22]. Recent studies have explored the use of large-scale pretrained Automatic Speech Recognition (ASR) models such as Whisper for feature extraction, demonstrating significant robustness improvements on English datasets [18]. However, low-resource languages remain critically underexplored, despite their linguistic diversity and the increasing accessibility of voice synthesis technologies.

Urdu represents a notable gap in current research. It is a morphologically rich and phonetically diverse Indo-Aryan language, characterised by aspirated and retroflex consonants, complex prosodic patterns, and accent variation [15]. These linguistic properties interact differently with synthesis and detection models compared to English. At the same time, the scarcity of large, labelled Urdu deepfake datasets limits model training and evaluation. Most benchmark corpora, including ASVspoof [12], FakeAVCeleb [13], and WaveFake [14], are English-centric. Only one publicly available Urdu deepfake audio dataset currently exists [11], and it has not been systematically used to evaluate modern architectures.

Prior work on Urdu speech technologies has primarily focused on ASR and speech synthesis, with limited emphasis on security applications. Deepfake detection studies have generally evaluated CNN and RNN architectures with handcrafted spectral features, achieving competitive performance on controlled datasets but often lacking robustness to distributional shifts or adversarial perturbations [19], [23], [24], [25]. Kawa et al. [18] showed that Whisper-based features can substantially improve deepfake detection accuracy and generalisation on English benchmarks, but their applicability to low-resource languages has not been investigated.

This paper addresses these gaps by presenting, to the best of our knowledge, the first structured benchmark of deepfake audio detection methods for Urdu. We evaluate representative architectures Light Convolutional Neural Networks (LCNN), CNN-LSTM with Attention, and Whisper Transformers combined with MFCC and LFCC front-end features. To mitigate data scarcity, we employ three dataset configurations (baseline, augmented, and extended) derived from an existing Urdu corpus [11] and enriched through controlled augmentation and additional recordings. We also incorporate robustness analyses under realistic signal perturbations such as additive noise, pitch shifts, and tempo variations.

The main contributions of this work are summarized as follows:

- We conduct the first systematic benchmarking of deepfake audio detection techniques for Urdu, providing

comparative insights across CNN, recurrent, and transformer architectures.

- We analyse how Urdu's linguistic characteristics influence model performance, complementing existing English-centric research.
- We introduce a structured dataset evaluation framework to assess generalisation and robustness under diverse signal conditions.
- We perform robustness and statistical analyses to evaluate model stability under realistic perturbations.

By situating Urdu within the broader landscape of multilingual deepfake detection, this study aims to inform the development of more inclusive and reliable speech forensics systems that address emerging security challenges across diverse linguistic contexts.

This paper is structured as follows: Section II discusses related work and previous methodologies. Section III presents the proposed models and approach. Section IV provides the results and analysis, followed by a detailed discussion in Section V. Finally, Section VI concludes the study with key findings and future directions.

## II. LITERATURE REVIEW

Early research on deepfake audio detection relied on traditional machine learning methods and handcrafted acoustic features. Techniques based on Mel-Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Perceptual Linear Prediction (PLP), and Constant Q Cepstral Coefficients (CQCC) have shown strong performance on benchmark datasets such as ASVspoof [12], FakeAVCeleb [13], and WaveFake [14]. These approaches achieved high detection accuracy in controlled settings, but their effectiveness declined when exposed to previously unseen synthesis methods or domain shifts [16], [24].

With the advancement of deep learning, researchers adopted spectrogram-based CNN and RNN architectures to automatically learn discriminative features from speech. Models such as VGG-16, MesoNet, LCNN, and CNN-LSTM hybrids have demonstrated improved performance compared to handcrafted features, particularly when trained on large English datasets [17], [19], [25], [26]. CNNs are effective in capturing local spectral patterns, while recurrent layers (e.g., LSTM, BLSTM) improve modelling of temporal dependencies. However, these models remain vulnerable to adversarial perturbations and often lack generalisation across languages and unseen attack types [23], [27].

To enhance robustness, multi-feature and feature fusion methods have been investigated. Combining handcrafted features with deep embeddings has led to improved detection rates, but often at the cost of higher computational complexity and potential overfitting [19], [26]. Lightweight CNN variants such as LCNN have been proposed to address real-time constraints, though sometimes with reduced accuracy on complex datasets [28].

A more recent direction involves leveraging self-supervised and transformer-based ASR models, such as

wav2vec 2.0 and Whisper, as powerful front-end feature extractors [18], [20], [21]. These models are trained on large multilingual corpora and capture rich phonetic and prosodic information. Kawa et al. [18] demonstrated that Whisper-based embeddings significantly reduce Equal Error Rates (EER) compared to conventional MFCC–CNN pipelines for English deepfake detection, highlighting their robustness to distributional shifts. Similar improvements have been observed with wav2vec-based models [20], [22], though cross-lingual evaluations remain limited.

Despite this progress, several gaps remain. First, nearly all studies focus on high-resource languages, primarily English, leaving low-resource languages underrepresented. Second, linguistic factors such as phonetic complexity, prosody, and accent variation which may affect synthesis and detection are rarely analysed systematically. Third, most evaluations rely on single datasets without cross-dataset or robustness testing, which limits the reliability of results in real-world conditions [23], [29]. Finally, the application of modern transformer-based models like Whisper has not been explored for deepfake detection in low-resource languages such as Urdu.

This study addresses these gaps by systematically benchmarking CNN, CNN–LSTM, and Whisper Transformer architectures for Urdu deepfake audio detection, using MFCC and LFCC features and evaluating robustness under diverse signal conditions.

### III. METHODOLOGY

This study employs a comprehensive methodology to develop an effective deepfake audio detection framework for the Urdu language as shown in the Figure 1. The approach begins with the construction of three distinct datasets Baseline, Augmented, and Extended, each designed to enhance model robustness and generalization. The Baseline dataset comprises original audio samples, whereas the Augmented dataset applies data augmentation techniques to introduce variability. The Extended dataset further enriches the training data by incorporating additional real-world audio samples, ensuring broader coverage and improved detection capability.

To classify deepfake and bonafide speech, two primary categories of deep learning architectures are explored:

- 1) **CNN-Based Architectures:** This includes Light Convolutional Neural Network (LCNN) [28] and CNN–LSTM with Attention, trained using two prominent acoustic feature extraction techniques: Mel-Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC). These architectures capture spatial dependencies through convolutional layers and temporal dependencies via recurrent layers, enabling the detection of deepfake artifacts.
- 2) **Transformer-Based Models:** This category comprises Whisper-tiny, Whisper-small, and Whisper-base, fine-tuned for Urdu deepfake detection. These pretrained ASR models leverage attention mechanisms to model

long-range dependencies in speech, allowing for improved deepfake identification.

A comparative analysis is conducted to assess the performance of these models in terms of accuracy, precision, recall, and Equal Error Rate (EER). The study further examines each model’s trainable parameters, evaluating trade-offs between computational efficiency and classification performance.

This section establishes a structured methodology for assessing the effectiveness and generalizability of the proposed deepfake detection models by systematically detailing dataset construction, model architectures, and evaluation metrics. The end-to-end pipeline from dataset loading to model evaluation is visually summarized in Fig. 1.

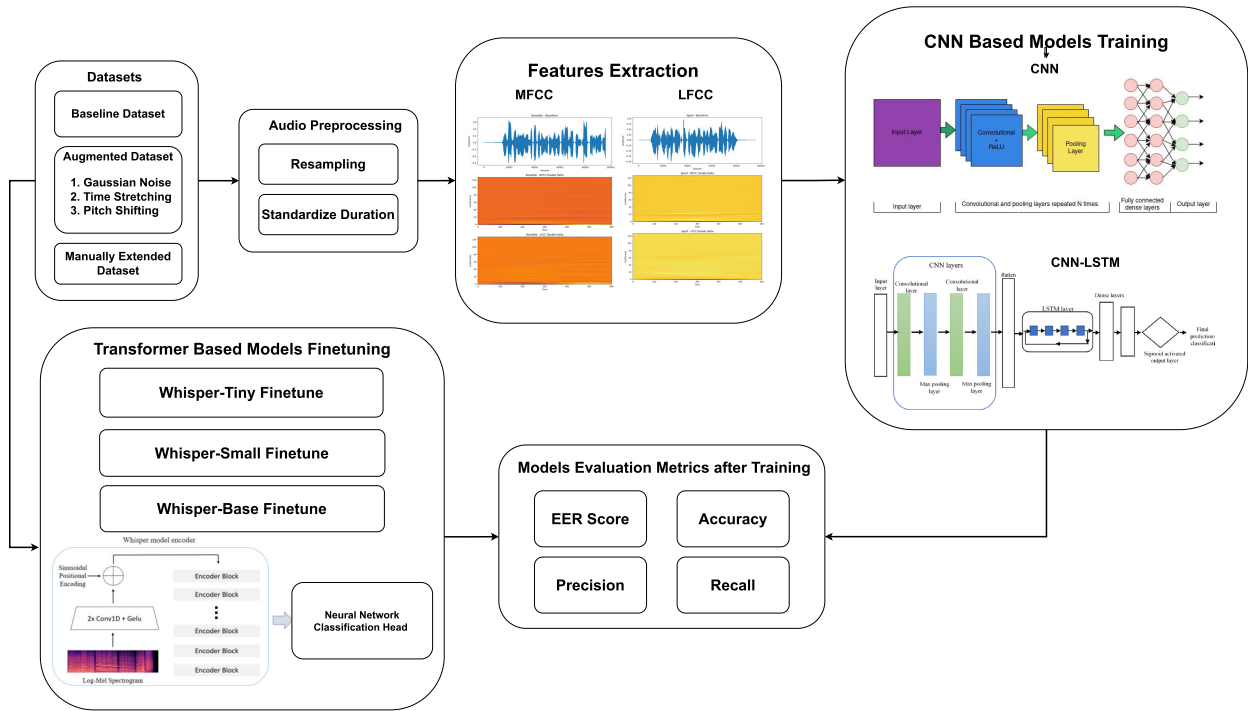
#### A. DATASET

This study employs the publicly available Urdu deepfake audio dataset introduced by Munir et al. [11], which includes both bona fide and synthetic speech samples from multiple male and female speakers with diverse regional accents. Synthetic audio was generated using state-of-the-art text-to-speech (TTS) and voice cloning systems to closely mimic genuine speakers. To systematically assess model performance under different data conditions, we constructed three dataset configurations:

- 1) **Baseline Dataset:** The original Urdu dataset without modification, consisting of 6,793 samples (3,397 bona fide and 3,396 spoofed), as shown in Table 1. This dataset serves as the primary benchmark. For all experiments, we applied a 70/15/15 stratified split, resulting in 4,755 training samples, 1,019 validation samples, and 1,019 test samples, preserving the class ratio across splits.
- 2) **Augmented Dataset:** Formed by applying three augmentation techniques Gaussian noise addition, pitch shifting, and time stretching to each Baseline sample. The augmentation parameters used were:
  - Gaussian noise:  $\sigma = 0.005$
  - Pitch shift:  $\pm 1.5$  semitones
  - Time stretch: factors randomly sampled from 0.9 to 1.1

Each original audio sample produced three augmented variants, expanding the dataset to 27,176 samples (13,590 bona fide and 13,586 spoofed) as shown in Table 1. The same 70/15/15 split was applied, resulting in 19,023 training samples, 4,076 validation samples, and 4,077 test samples.

- 3) **Extended Dataset:** Constructed by adding supplementary Urdu speech recordings from publicly available corpora, balanced across gender and accent groups. Synthetic counterparts were generated using the same TTS pipelines as the original dataset. This expanded the dataset to 20,383 samples (10,877 bona fide and 9,506 spoofed) as shown in Table 1. Using the same stratified 70/15/15 split, the dataset was partitioned into 14,268 training samples, 3,058 validation samples,



**FIGURE 1.** The figure illustrates the overview of the proposed methodology, starting with the dataset as input. Augmentation techniques are applied only to the baseline dataset, while the extended dataset is passed directly to the preprocessing step. Features, such as MFCC or LFCC, are then extracted from the preprocessed data. The next step involves training two categories of models: CNN-based models and Transformer-based models. Finally, the models are evaluated using accuracy, precision, recall, F1 score, and EER score.

and 3,057 test samples. Unlike the Augmented dataset, the Extended dataset introduces entirely new real and synthetic audio, increasing linguistic and acoustic diversity.

**TABLE 1.** Statistics of the baseline, augmented, and extended datasets, including train/validation/test splits.

Category	Baseline	Augmented	Extended
Total Samples	6,793	27,176	20,383
Bonafide Samples	3,397	13,590	10,877
Spoofed Samples	3,396	13,586	9,506
Train Samples	4,755	19,023	14,268
Validation Samples	1,019	4,076	3,058
Test Samples	1,019	4,077	3,057

**B. DATA PREPROCESSING**

Effective deepfake audio detection requires consistent and standardized data processing. This step ensures uniformity across the dataset, facilitating efficient batch processing and enhancing model performance.

**1) SAMPLING RATE**

The sampling rate defines the number of audio samples captured per second, playing a crucial role in speech processing. This study employs a sampling rate of 16,000 Hz (16 kHz), which is widely used in speech-based deep learning applications. This rate balances audio quality and computational efficiency, as higher rates (e.g., 44.1 kHz)

increase data size without notable benefits, whereas lower rates may compromise voice clarity [30].

**2) FIX AUDIO DURATION**

To maintain consistency in input length, all audio samples are fixed to a maximum duration of 4 seconds. If an audio sample is greater than the defined limit then it would be truncated, and if shorter than zero-padding is applied. This step ensures that all the input samples have the same dimension, which is important for efficient batch processing in deep learning models. This observation indicated that 4 seconds is sufficient to capture meaningful speech features whereas minimizing computational overhead.

**C. DATA AUGMENTATION**

After the successful preprocessing, the data augmentation techniques are applied to bring the generalizability and variations in the data achieve the robustness in the deepfake audio detection.

Augmentation is essential as it introduces variability in speech characteristics, enabling the model to adapt to real-world distortions and unseen scenarios. This study employs three key augmentation methods: Gaussian noise injection, pitch shifting, and time stretching. Each technique plays a distinct role in increasing dataset diversity, ensuring that the model captures abstract speech representations rather than relying on speaker-specific traits or synthetic artefacts.

The Gaussian noise augmentation in this study adopted a fixed standard deviation ( $\sigma = 0.005$ ), which provides moderate signal perturbation without overwhelming low-energy phonetic cues. However, fixed-noise augmentation captures only a narrow range of real-world recording conditions. In practice, background noise varies substantially across microphones, environments, and transmission channels. Using a single  $\sigma$  value may therefore limit robustness. Future extensions should incorporate variable-noise schedules (e.g.,  $\sigma \in 0.001\text{--}0.02$ ) to simulate diverse acoustic conditions and improve generalisation under mismatched noise levels.

### 1) GAUSSIAN NOISE

Gaussian noise injection introduces random noise into the audio waveform to simulate real-world environmental interference. The added noise follows a Gaussian (normal) distribution and is mathematically defined in Equation 1:

$$x_{\text{noisy}}[n] = x[n] + N(0, \sigma^2) \quad (1)$$

where:

- $x[n]$  represents the original audio signal,
- $N(0, \sigma^2)$  is Gaussian noise with zero mean and variance  $\sigma^2$ ,
- $x_{\text{noisy}}[n]$  is the resulting noisy audio signal.

The standard deviation  $\sigma$  is set to a small value (e.g., 0.005) to maintain speech intelligibility whereas preventing excessive distortion. This augmentation is particularly beneficial in replicating background noise conditions, such as wind, traffic, and microphone artifacts, ensuring that models become robust to real-world distortions. Additionally, noise injection aids in preventing overfitting, as models trained solely on clean audio may struggle to generalize to real-world environments.

### 2) PITCH SHIFTING

Pitch shifting modifies the fundamental frequency of an audio signal whereas preserving its temporal structure. This transformation is applied within a semitone range of  $\pm 1.5$ , introducing speaker variability without altering speech intelligibility. The process involves:

- 1) Time-stretching the audio.
- 2) Resampling to maintain the original duration whereas shifting the pitch.

This augmentation is essential as speech pitch naturally varies among different speakers and across emotional states. By incorporating pitch variations, the model learns deeper speech representations instead of relying on speaker-specific frequency characteristics. Without this augmentation, models risk overfitting to individual speaker tones, reducing their ability to detect unseen deepfake speech. Furthermore, deepfake synthesis often introduces unnatural pitch fluctuations, making pitch variability a valuable feature for robust detection.

### 3) TIME STRETCHING

Time stretching modifies the duration of an audio signal without affecting its pitch. This is implemented using a phase vocoder algorithm, which:

- Divides the signal into overlapping frames.
- Adjusts the phase to maintain consistency.
- Reassembles frames to produce a stretched version of the original audio.

The time stretch factor is chosen randomly within a range of 0.9 to 1.1, meaning that audio can be sped up or slowed down by up to 10%. This augmentation is critical for simulating natural variations in speech rate, which occur due to factors such as emotional state, cognitive load, and regional dialects.

By training on time-stretched audio, the model avoids overfitting to fixed speech tempos, improving its ability to generalize to diverse speaking styles. This is particularly useful in deepfake detection, as synthetic speech often exhibits rigid and unnatural timing patterns, making this augmentation a crucial component for enhancing model robustness.

## D. FEATURE EXTRACTION

Feature extraction is a fundamental step in speech processing that transforms raw waveforms into numerical representations for classification. This study employs two widely used features: Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Frequency Cepstral Coefficients (LFCCs). These features capture key spectral characteristics of speech, which are essential for distinguishing between real and deepfake audio.

### 1) MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

The Mel-Frequency Cepstral Coefficients (MFCCs) are commonly applied in speech processing as they can resemble the human hearing process because of their focus on the lower-frequency range. The method is especially efficient in detecting deepfakes because this type of feature extraction is capable of capturing key frequency-related traits of speech, which are needed to distinguish between natural and synthetic audio. The MFCC extraction algorithm starts by computing 128 coefficients which are further reduced by first-order (delta) and second-order (two times the derivative, double delta) derivatives to add time variation. This will enable the model to learn unnatural patterns that are added through deepfake synthesis methods.

It begins with pre-emphasis, the application of a high-pass filter to boost high-frequency content that would otherwise be oppressed when one is producing speech. This enhances the signal's spectral details making subtle distortions introduced by synthetic speech more detectable. Next, the audio signal is framed and windowed, where it is divided into overlapping 25 ms segments with a 10 ms overlap. Each frame is then multiplied by a Hamming window, which helps in reducing

spectral leakage by tapering the signal edges, ensuring smoother frequency transitions.

Once the signal is framed, it undergoes a Fast Fourier Transform (FFT) to convert it into the frequency domain. Since human auditory perception is nonlinear, the transformed spectrum is passed through a Mel filterbank, which maps the spectrum to the Mel scale, ensuring that perceptually significant frequency bands are emphasized. After this transformation, the logarithm of the filterbank energies is computed, followed by the Discrete Cosine Transform (DCT). This step helps in decorrelating the coefficients, reducing redundancy and allowing for a more compact feature representation. Typically, only the first 13 coefficients are retained as they encapsulate the most relevant speech information.

To further enhance the model’s capability in temporal pattern recognition, delta and double delta features are computed. These features capture how MFCC values change over time, enabling the model to differentiate between naturally occurring speech variations and artificially generated speech artifacts. The final feature vector comprises 39 dimensions, consisting of 13 static MFCCs, 13 delta MFCCs, and 13 double delta MFCCs, ensuring a rich temporal representation crucial for deepfake detection. The entire MFCC extraction process is illustrated in Fig. 2.

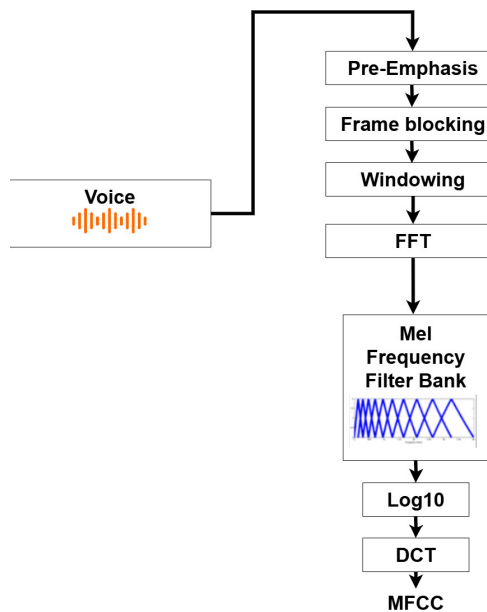


FIGURE 2. MFCC feature extraction.

## 2) LINEAR FREQUENCY CEPSTRAL COEFFICIENTS (LFCC)

Linear Frequency Cepstral Coefficients (LFCCs) follow a similar extraction process as MFCCs but differ in how they represent the spectral content of speech. Unlike MFCCs, which apply Mel scaling to emphasize lower frequencies, LFCCs operate in the linear frequency domain, ensuring uniform spectral energy distribution across all frequency bands. This characteristic makes LFCCs particularly useful

for forensic applications, where a more detailed spectral analysis is required.

In the LFCC extraction process, the audio signal undergoes pre-emphasis, framing, and windowing, similar to MFCCs, ensuring consistency in feature extraction. However, instead of using the Mel scale, a linearly spaced filterbank is applied, maintaining equal spectral resolution across all frequencies. This allows more sensitivity to high-frequency content, useful when synthetic speech artifacts lie outside the audible range of the human ear.

Similar to MFCCs, LFCCs too have delta and double delta features that are appended to the original LFCCs to increase temporal resolution. This ensures that the model captures both spectral and dynamic variations in speech, allowing for improved deepfake detection accuracy. The extraction process for LFCCs is depicted in Fig. 3.

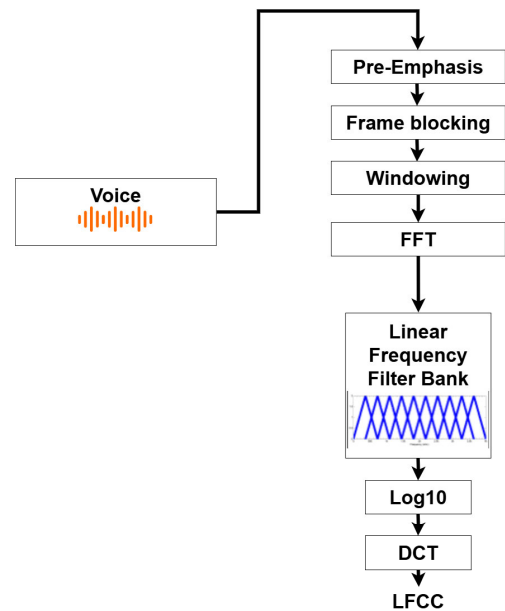


FIGURE 3. LFCC feature extraction.

## E. CLASSIFICATION MODELS

Detecting deepfake audio is a complex task that requires a robust feature extraction and classification pipeline capable of distinguishing synthetically generated speech from bonafide human recordings. Given the advancements in AI-driven speech synthesis, deepfake audio presents challenges such as imperceptible artifacts, prosodic inconsistencies, and subtle frequency distortions that are often difficult to detect using traditional methods. To address these challenges, this study explores three deep learning architectures optimized for deepfake detection: Light Convolutional Neural Network (LCNN) with Bi-directional Long Short-Term Memory (BLSTM), CNN-LSTM with Attention, and Whisper-based Transformer models. Each model is designed to leverage both spatial and temporal dependencies in speech whereas maintaining computational efficiency.

### 1) LIGHT CONVOLUTIONAL NEURAL NETWORK (LCNN) WITH BI-DIRECTIONAL LSTM

The LCNN-BLSTM model processes spectrogram-based audio inputs, capturing spatial features through convolutional layers and temporal dependencies via recurrent layers. The convolutional front-end consists of six layers that progressively refine extracted features. The first convolutional layer applies 64 filters of size  $5 \times 5$  with Max-Feature-Map (MFM) activation, enhancing feature selection whereas reducing redundancy. Batch normalization ensures stable activation distributions across training batches. Deeper layers employ 96 and 128 filters with smaller kernel sizes ( $3 \times 3$ ,  $1 \times 1$ ) for hierarchical feature extraction. Max-pooling layers progressively reduce dimensionality whereas preserving essential speech characteristics.

Following feature extraction, the BLSTM module processes the spectrogram-derived sequences. Unlike conventional recurrent models, BLSTM networks capture bidirectional dependencies, allowing the model to interpret both preceding and succeeding contextual information in speech. Two BLSTM layers, each with 64 hidden units per direction, encode sequential dependencies. The final global temporal averaging layer aggregates sequential representations into a fixed-length feature vector, which is then mapped to a binary classification decision (bonafide vs. deepfake) through a fully connected layer with a softmax activation.

### 2) CNN-LSTM WITH ATTENTION MECHANISM

Whereas the LCNN-BLSTM model effectively captures both spatial and temporal dependencies, it does not differentiate between important and unimportant time steps in sequential representations. To address this limitation, a CNN-LSTM with Attention model is introduced, integrating feature extraction, sequential modeling, and adaptive attention weighting to refine deepfake detection performance. The convolutional front-end is similar to LCNN, extracting spatial features from spectrograms by detecting pitch variations, formant structures, and frequency artifacts. However, instead of relying solely on BLSTM encoding, the extracted features are flattened into sequential representations and processed using Bi-directional LSTM layers, which capture time-dependent variations in speech.

A key enhancement in this model is the adaptive attention mechanism, which assigns dynamic importance scores to each time step in the sequential representation. Traditional recurrent models treat all time steps equally, which may lead to information redundancy and ineffective learning from crucial deepfake patterns. The attention module computes contextual importance scores using a softmax-weighted alignment mechanism, allowing the model to focus on salient speech regions that exhibit synthetic artifacts or prosodic anomalies. This selective weighting enhances classification performance by enabling the network to attend to critical speech distortions introduced during deepfake generation.

The final fully connected classification layer takes the attention-weighted representations and produces a binary classification score.

CNN-based models were selected because they are highly effective at identifying short-duration spectral anomalies that commonly appear in Urdu synthetic speech, including irregular aspiration noise, incomplete retroflex burst characteristics, and unnatural nasalization transitions. These artefacts are localized within specific time–frequency regions, making convolutional filters well-suited for capturing them. The addition of temporal modeling components such as LSTMs or attention enables the network to track longer-range prosodic patterns that are characteristic of Urdu speech, providing an architecture aligned with the language’s spectro-temporal structure.

### 3) WHISPER-BASED TRANSFORMER FOR DEEPPAKE DETECTION

Whereas convolutional and recurrent networks provide effective local feature extraction and temporal modeling, they are limited in capturing long-range dependencies present in speech signals. To overcome this limitation, we introduce a Whisper-based Transformer model, which leverages self-attention mechanisms to learn hierarchical representations of speech. Unlike CNN-LSTM architectures, which rely on sequential processing, Transformers process entire input sequences in parallel, allowing for efficient long-range feature dependencies to be captured.

The Whisper model was originally designed for automatic speech recognition (ASR) and trained on massive multilingual datasets, making it a powerful feature extractor for deepfake detection. However, for deepfake classification, we remove the decoder component and repurpose Whisper’s encoder as a high-dimensional speech representation extractor. The raw audio waveform is first tokenized using Whisper’s internal processor and converted into log-Mel spectrogram representations, which are then fed into the Transformer encoder. This encoder maps the input to high-dimensional embeddings that encapsulate phonetic structures, speaker identity, and spectral distortions. Unlike conventional Whisper applications, which use these embeddings for speech-to-text transcription, we introduce a binary classification head that processes these embeddings and predicts bonafide or deepfake audio.

To optimize Whisper for deepfake detection, the model undergoes supervised fine-tuning using a dataset of real and synthetic Urdu speech samples. The fine-tuning process involves unfreezing selected encoder layers, allowing the model to learn deepfake-specific audio patterns whereas retaining the generalized speech representations from its pretraining phase. Training is conducted using the Cross-Entropy Loss function, with the AdamW optimizer applied to ensure efficient weight updates and regularization. The model is then evaluated on multiple configurations, including Whisper-tiny, Whisper-small, and Whisper-base, with final

model selection based on the trade-off between detection accuracy and computational efficiency.

The system utilizes Whisper’s encoder-only pathway, with log-Mel features generated using the model’s native 25 ms window and 10 ms hop size. During fine-tuning, the lower six encoder layers were frozen to retain multilingual pretraining priors, while the remaining layers and projection head were updated. Embeddings were mean-pooled across time before classification. Mixed-precision training, gradient checkpointing, and a two-epoch warm-up schedule were employed to stabilize fine-tuning on Urdu deepfake data.

#### F. TRAINING HYPERPARAMETERS

All CNN-based models (LCNN and CNN-LSTM-Attention) were trained using a 70/30 train–test split. We optimized these architectures with the Adam optimizer (learning rate  $1 \times 10^{-4}$ , batch size 16) for 30 epochs. The LCNN employs max-pooling for spatial downsampling, batch normalization after several convolutional layers, and a dropout layer with  $p = 0.5$  in the final convolutional block for regularization. For the Whisper-based model, the data were divided into training, validation, and test sets with a 70/15/15 ratio. Only the last three layers of Whisper-base were fine-tuned, using the AdamW optimizer (learning rate  $1 \times 10^{-5}$ , weight decay 0.01, batch size 16) for 30 epochs. All models were trained using cross-entropy loss.

#### G. EXPERIMENTAL SETUP AND COMPLEXITY

The Tables 2 and 3 present model specifications for various architectures. They report the number of parameters, FLOPs, and average inference time per forward pass. All experiments were conducted on a system equipped with an NVIDIA GeForce RTX A2000 GPU (12GB VRAM), 32GB RAM, and fifteen 512GB NVMe SSDs configured for high-speed storage. The FLOPs and parameter counts were obtained using the thop profiling library, while inference time was averaged over 100 runs in evaluation mode to minimize variance. These results provide a comparative reference for model complexity and runtime efficiency, supporting informed architectural choices based on available computational resources.

### IV. EXPERIMENTS AND RESULTS

This section presents a comprehensive evaluation of the proposed deepfake audio detection framework for the Urdu language. The experiments examine the impact of feature extraction methods, model architectures, and data diversity on detection performance. Two model families are assessed: (i) CNN-based models, including LCNN and CNN-LSTM with attention, using MFCC and LFCC features; and (ii) Transformer-based models based on Whisper variants of different capacities. The evaluation considers three dataset configurations baseline, augmented, and extended—to assess generalisation under progressively richer training conditions. Performance is measured in terms of classification accuracy and Equal Error Rate (EER), which together provide complementary perspectives on discriminative power

and calibration. The results are presented in four parts: Section IV-A analyses CNN-based models, Section IV-B examines Transformer-based models, Section IV-C provides cross-model comparisons and contrasts with prior work, and Section IV-D summarises key findings.

#### A. PERFORMANCE OF CNN-BASED MODELS

Fig. 7 presents a comparative analysis of CNN-based architectures, focusing on classification accuracy and Equal Error Rate (EER) across different feature extraction techniques (MFCC, LFCC) and dataset configurations (baseline, augmented, extended). The results highlight the critical interaction between feature design, model architecture, and data diversity in determining detection performance.

For MFCC features, the baseline CNN model achieved 79 % accuracy with an EER of 20.88 %. While MFCCs are well established for capturing the perceptually relevant spectral characteristics of speech, these baseline results indicate clear room for improvement in discriminative power and calibration.

Introducing temporal modeling and attention significantly improved performance. The CNN-LSTM with attention model using MFCC features reached 82 % accuracy and an EER of 17.93 %. The LSTM layers effectively captured long-term temporal dependencies, while the attention mechanism emphasised the most informative regions of the signal, leading to better classification of bona fide and synthetic speech.

LCNN under baseline conditions performed markedly worse, with 58 % accuracy and a high EER of 42.20 %. Although LCNNs are computationally lightweight and advantageous for real-time applications, their limited representational capacity makes them less effective for complex speech signals. Combining LCNN with LSTM and attention improved accuracy to 74 % and reduced EER to 25.09 %, but performance still lagged behind MFCC-based CNN-LSTM models, reflecting LCNN’s architectural constraints.

To analyze the learned representations, Fig. 4 shows a t-SNE visualization of MFCC-based LCNN embeddings. The projection exhibits clear separation between bona fide and spoofed utterances, indicating effective discriminative learning. Misclassified samples, highlighted in the embedding space, are sparsely distributed near the class boundary, suggesting acoustic overlap between real and synthetic speech. Analysis of the corresponding misclassified spectrograms (Fig. 5 and Fig. 6) identifies three main error sources: limited temporal context in short utterances, spectral distortions caused by overlapping speech and background noise, and strong regional accent variations in Urdu (e.g., aspirated stops, retroflex consonants, and breathy phonation) that partially overlap with synthesis artifacts. For clarity, results are shown only for the best-performing LCNN model, as similar trends were observed across other architectures.

LFCC features exhibited a similar pattern. Baseline CNN with LFCC achieved 74 % accuracy and 25.90 % EER. Adding CNN-LSTM with attention improved this

TABLE 2. Model metrics for MFCC and LFCC features.

Model	Feature Type	Params (K)	Params (M)	FLOPs (M)	FLOPs (G)	Avg. Time (ms)
CNN LSTM Attention	MFCC	357.95	0.36	729.40	0.73	3.24
LCNN	MFCC	343.43	0.34	920.37	0.92	1.51
CNN LSTM Attention	LFCC	357.95	0.36	729.40	0.73	3.18
LCNN	LFCC	343.43	0.34	920.37	0.92	1.11

TABLE 3. Whisper variants models specifications.

Model	Total Params (M)	Trainable Params (K)	Size (MB)	FLOPs (G)	Inference Time (ms)
Tiny	8.31	99.84	31.69	11.73	6.75
Base	20.72	132.87	79.05	30.10	2.20
Small	88.35	198.91	337.03	131.02	10.65

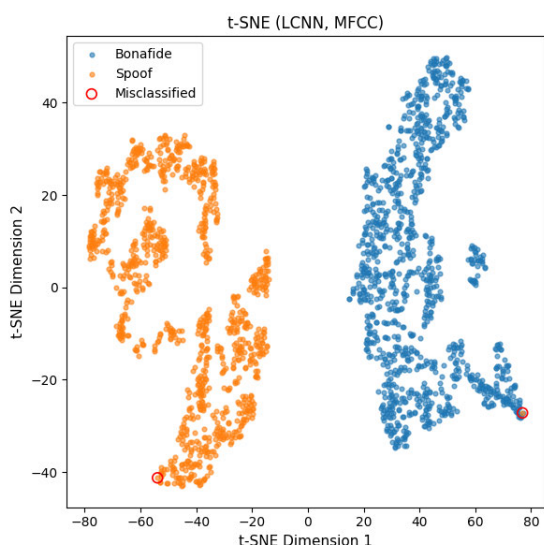


FIGURE 4. t-SNE visualization of LCNN embeddings using MFCC features. Bona fide and spoofed utterances form two well-separated clusters, while misclassified samples (highlighted in red) appear near the class boundary, indicating overlapping acoustic characteristics.

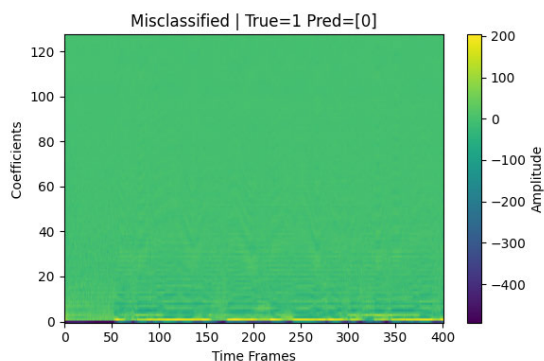


FIGURE 5. Time-frequency representation of a misclassified bona fide utterance predicted as spoof, exhibiting near-natural spectral structure.

to 81 % accuracy and 19.38 % EER. While LFCCs can represent linear frequency characteristics effectively, they were consistently outperformed by MFCCs, likely because

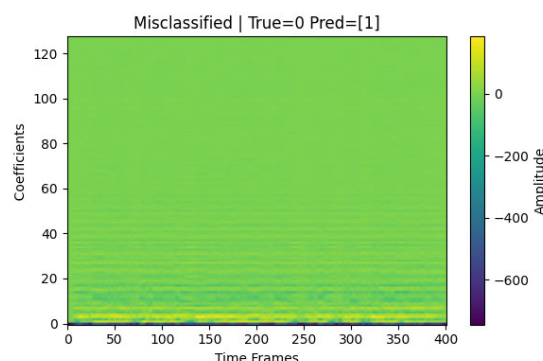


FIGURE 6. Time-frequency representation of a misclassified spoof utterance predicted as bona fide, showing noise contamination and flattened spectral dynamics.

Mel-scaled features better expose subtle synthesis artefacts present in Urdu deepfake audio.

Urdu’s phonetic system contains dense formant transitions, breathy phonation, and aspirated consonants that create strong low- to mid-frequency spectral cues. MFCCs, which emphasize perceptually relevant mel-scaled regions, capture these phenomena more effectively than LFCCs. LFCCs distribute resolution uniformly across the spectrum, which can dilute energy in precisely those bands where Urdu-specific phonetic contrasts reside, reducing separability between bona fide and synthetic speech. Moreover, many Urdu deepfake synthesis systems fail to model breathy voice onset and retroflex bursts accurately—artefacts that MFCCs detect more reliably, explaining their consistently superior performance.

Data augmentation produced further improvements. Using MFCC features with augmented data increased accuracy to 81 % and reduced EER to 19.38 %, highlighting the benefits of introducing variability through noise injection, pitch shifting, and tempo modifications. Such augmentation improves generalisation by exposing models to a broader range of acoustic conditions and speaker variations.

The highest CNN-based performance was achieved by the CNN-LSTM with attention model using MFCC features on the extended dataset, which reached an accuracy of 98 % and an exceptionally low EER of 0.52 %. This improvement reflects the combined benefits of richer data coverage and architectural enhancements such as deeper networks and stronger regularization, which enable the model to capture more complex acoustic patterns. The remarkably low EER demonstrates the model's ability to generalize effectively to unseen data, underscoring the strength of advanced CNN architectures for large and heterogeneous speech datasets.

By contrast, the CNN-LSTM with attention model using LFCC features on the extended dataset achieved a lower accuracy of 75 % with an EER of 25.08 %. Although LFCCs can capture detailed linear frequency information, these results indicate that they are less effective than MFCCs in representing synthesis artefacts in Urdu deepfake audio. The higher error rates suggest greater difficulty in managing both false acceptances and false rejections relative to MFCC-based models.

Overall, these results highlight the critical role of feature extraction, model architecture, and dataset diversity in determining CNN-based model performance. MFCC features consistently outperform LFCC features, particularly when combined with temporal modelling and attention mechanisms. Data augmentation and extension further enhance model robustness by broadening the range of acoustic and speaker variations encountered during training. Collectively, these factors enable CNN-based architectures to achieve strong performance on challenging speech classification tasks, especially in scenarios where temporal dynamics and feature richness are central to detection accuracy.

## B. PERFORMANCE OF TRANSFORMER-BASED MODELS

The performance of Transformer-based architectures was evaluated using three Whisper variants Whisper-tiny, Whisper-small, and Whisper-base under baseline and extended dataset configurations. Accuracy and Equal Error Rate (EER) were used as evaluation metrics, enabling a direct comparison with CNN-based approaches. The results are summarised in Fig. 8.

Under baseline conditions, Whisper-tiny achieved an accuracy of 82 % and an EER of 18.20 %, while Whisper-small slightly improved to 84 % accuracy and 15.88 % EER. This performance difference reflects the benefit of increased model capacity: larger models are generally able to learn more complex representations, leading to improved classification and calibration. In contrast, Whisper-base achieved 75 % accuracy and a substantially higher EER of 25.15 %. This indicates that simply scaling up model size does not guarantee better performance in data-sparse settings such as Urdu; larger models may overfit or fail to generalise effectively without sufficient training diversity.

Extending the models resulted in substantial performance gains across all three Whisper variants. The extended Whisper-tiny model reached 93 % accuracy with a

significantly reduced EER of 6.70 %, demonstrating that additional data diversity and fine-tuning enable even small Transformer models to generalise more effectively. Whisper-small achieved the best overall results, with 99 % accuracy and an exceptionally low EER of 0.50 %. Similarly, Whisper-base reached 99 % accuracy with an EER of 0.58 %. These improvements highlight the strong capacity of Transformer encoders to model long-range temporal dependencies and detect subtle synthesis artefacts when fine-tuned on richer training data.

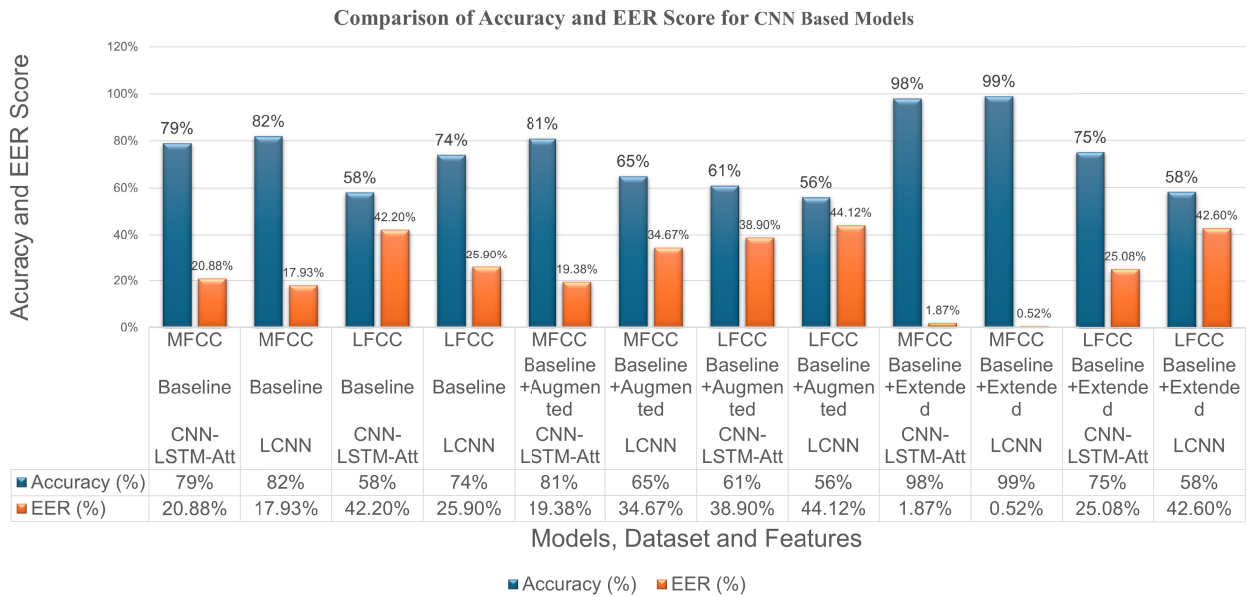
The consistent improvements observed across all Whisper variants underscore two important findings. First, model extension through data augmentation and domain-specific fine-tuning is crucial for achieving state-of-the-art performance, regardless of the base model size. Second, Transformer architectures, particularly Whisper-small and Whisper-base, exhibit exceptional generalisation capabilities when appropriately adapted, achieving near-perfect classification accuracy and minimal error rates.

## C. CROSS-MODEL COMPARISON AND PRIOR WORK

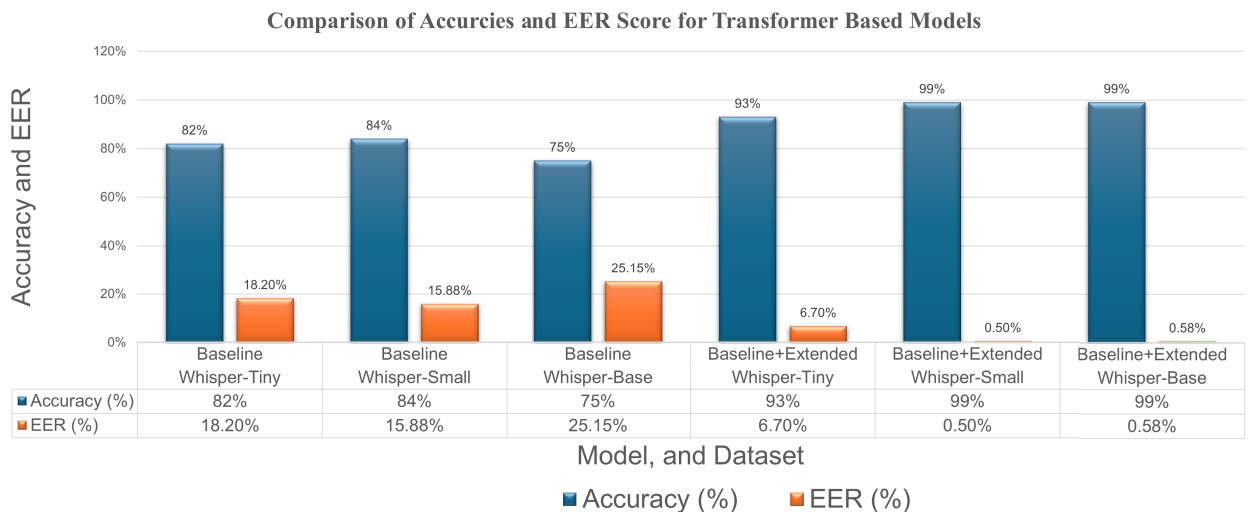
Table 4 summarises EER values across different models, features, and datasets, and compares them to prior work. The weakest performance was observed on the DeepFake-In-The-Wild dataset using LCNN with LFCC features, with an EER of 77.56 %. Using MFCC features with LCNN on the same dataset reduced the EER to 67.62 %, but results remained suboptimal. In contrast, the Urdu dataset yielded markedly lower error rates, with LFCC + LCNN achieving 25.90 % EER and CNN-LSTM with attention achieving further reductions. Whisper-based Transformer models, particularly Whisper-small and Whisper-tiny, demonstrated the best performance, achieving EERs as low as 0.50 % in extended configurations. This comparison illustrates the significant advantages of modern architectures and richer feature representations for deepfake detection in low-resource languages.

When compared with previous work by Kawa et al. [18], the differences are striking. Their study reported EERs of 67.62 % (LFCC + LCNN) and 32.90 % (MFCC + LCNN) on the DeepFake-In-The-Wild dataset. In contrast, the proposed models achieve substantially lower EERs on the Urdu dataset, with Whisper-small reaching 0.50 %. These results demonstrate the effectiveness of combining hybrid temporal architectures (e.g., CNN-LSTM with attention) and large multilingual Transformer encoders for capturing synthesis artefacts that traditional CNN-based approaches miss.

Overall, these findings highlight the decisive impact of both model architecture and training strategy. Baseline Whisper models already outperform many traditional CNN-based methods, and extended versions deliver near-perfect performance. This underscores the suitability of Transformer architectures for deepfake audio detection in low-resource settings and demonstrates clear improvements over prior state-of-the-art methods.



**FIGURE 7.** The figure shows the accuracy and EER score for the LCNN, CNN-LSTM with Attention, using both the MFCC and LFCC features and all three datasets.



**FIGURE 8.** Comparison of accuracy and EER for transformer-based models.

**D. COMPUTATIONAL EFFICIENCY ANALYSIS**

The comparative results reveal a clear trade-off between accuracy and computational cost. LCNN and CNN-LSTM models offer significantly lower FLOPs and inference latency (1.1–3.2 ms per sample) and are therefore suited for embedded forensic applications or real-time detection on edge hardware. However, their reduced representational capacity limits performance under noisy or cross-domain conditions. Whisper-based models, especially Whisper-small and Whisper-base, require substantially larger compute budgets—up to 131 GFLOPs and >300 MB of memory—but provide state-of-the-art accuracy and robustness. Whisper-tiny presents a balanced compromise, offering strong generalization with moderately increased

latency. For real-world deployment in telecom, banking, and authentication pipelines, this trade-off suggests a cascaded hybrid approach: lightweight CNNs for real-time triage and Whisper-based models for high-assurance verification.

**V. DISCUSSION**

The results provide clear evidence of the effectiveness of transformer-based models for deepfake audio detection in low-resource languages. Whisper consistently outperformed CNN and CNN-LSTM architectures across all datasets, achieving the lowest EER and highest AUC with minimal variance. This advantage reflects the ability of large-scale multilingual pretraining to capture linguistic and prosodic cues beyond what handcrafted spectral features provide.

These findings are consistent with prior English-language studies [18], [20], [21] and extend them to Urdu, a morphologically rich and phonetically diverse language.

Urdu exhibits several phonetic and prosodic properties that directly influence deepfake artefacts and detection performance. Its four-way stop system—voiced, voiceless, aspirated, and breathy-voiced—introduces greater variability in aspiration noise and low-frequency energy patterns compared to English. The extensive retroflex inventory produces distinct burst signatures and formant transitions that many synthesis models fail to reproduce accurately. Additionally, phonemic nasalization and breathy phonation generate mid-frequency spectral cues that are challenging for neural vocoders to synthesize. Urdu also displays longer vowel durations, richer coarticulation, and more flexible intonation contours than English. These characteristics contribute to unique spectro-temporal deviations in synthetic Urdu speech and therefore motivate specialized modeling approaches that differ from English-centric deepfake detection studies.

While the models were trained exclusively on Urdu data, the phonetic structure of Urdu offers important insights into cross-lingual generalization. Urdu contains a richer inventory of aspirated stops, retroflex consonants, and breathy-voiced segments compared to English and Hindi. These phonetic categories introduce greater within-speaker spectral variability, particularly in the low-mid frequency bands where MFCCs operate, which partially explains why MFCC-based architectures generalize better than LFCC-based ones. English, by contrast, exhibits fewer retroflex and aspirated contrasts, allowing synthetic models to produce smoother spectral contours, which often results in distinct artefact patterns learned by detectors. Hindi shares some phonetic similarities with Urdu but lacks Urdu's extensive prosodic variation and nasalization patterns, which further influences model robustness. These differences highlight that deepfake detectors trained on English-centric artefacts may not generalize well to Urdu, underscoring the need for language-specific or multilingual training pipelines.

Data diversity emerged as a critical factor. Augmentation and dataset extension significantly improved the performance of CNN-based models, reducing EER by up to 0.5%. This is particularly relevant for Urdu, where accent and prosodic variation are pronounced [15]. Whisper showed smaller but stable gains, indicating that its pretrained representations already encode substantial variability.

Robustness evaluations revealed distinct model vulnerabilities. CNN-based architectures were more affected by noise and pitch shifts, which disrupt MFCC and LFCC representations. Whisper remained stable under perturbations, suggesting that transformer embeddings are less reliant on low-level spectral regularities and better reflect higher-level linguistic structure. This robustness is advantageous for real-world forensic scenarios, where recordings may be noisy or partially manipulated.

Whisper's encoder relies on log-Mel spectrograms that preserve perceptually weighted spectral energies, which

naturally align with MFCCs' mel-scaled filterbanks [31]. This alignment means that MFCCs provide a feature space whose distribution complements the encoder's pretraining priors, enabling smoother optimization during fine-tuning and more stable convergence. In contrast, LFCCs emphasize linear spectral resolution—including high-frequency regions where transformers apply less attention density—leading to weaker representational synergy. As a result, MFCC-driven CNN-LSTM models benefit from higher sensitivity to prosodic and formant-level distortions that Whisper embeddings also encode.

Transformer embeddings already encode a rich mixture of spectral and temporal dependencies, raising the question of whether cepstral features become redundant. Our empirical results suggest partial redundancy but also meaningful complementarity. MFCCs capture fine-grained short-time spectral envelopes and emphasize harmonic-to-noise ratios—information that complements Whisper's distributed attention patterns, which focus more on longer-range prosody and phonotactics. Some spectral overlap exists, especially in mel-scaled regions, but the redundancies appear beneficial: overlapping cues reinforce artefact detection and reduce false negatives for subtle synthesis artefacts. LFCCs, with their uniform high-frequency emphasis, introduce less complementary information, explaining their overall weaker performance in Urdu detection.

Cross-dataset evaluations highlighted the superior generalisation of Whisper. Although performance declined when transferring between Urdu and English datasets, Whisper maintained lower EER than CNN-based models, underscoring the potential of multilingual transformers for cross-lingual detection. Nonetheless, the observed performance gap indicates that multilingual deepfake detection remains challenging.

A breakdown of model decisions indicates that different artefact classes contribute unequally to detection accuracy. Formant-shape distortions around retroflex and aspirated consonants showed the strongest discriminative impact, particularly for MFCC-driven CNN-LSTM models, which are sensitive to short-time spectral envelope deviations [11].

Whisper-based models relied more on prosodic irregularities, including unnatural timing, flattened intonation contours, and inconsistent breathiness, reflecting the transformer's broader contextual window [18], [31]. High-frequency noise artefacts contributed least to model decisions, especially for Urdu, where mid-band phonetic cues dominate the acoustic space. This hierarchy of artefact sensitivity highlights that Urdu deepfake detection is primarily driven by mid-frequency spectral-prosodic deviations rather than high-frequency GAN artefacts commonly exploited in English-centric systems [12].

Different model families exhibit distinct sensitivities to acoustic distortions. CNN-based models using MFCC and LFCC features are more affected by noise, pitch, and tempo variations due to their reliance on short-time spectral representations. In contrast, Whisper-based transformer models

**TABLE 4.** EER comparison of different models, features, and datasets with previous research work.

Reference	Dataset	Features	Model	EER (%)
[18]	DeepFake-In-The-Wild	LFCC	LCNN	77.56
[18]	DeepFake-In-The-Wild	MFCC	LCNN	67.62
[18]	DeepFake-In-The-Wild	Whisper	LCNN	32.90
Proposed Work	Urdu Dataset	LFCC	LCNN	25.90
Proposed Work	Urdu Dataset	MFCC	LCNN	0.52
Proposed Work	Urdu Dataset	MFCC	CNN-LSTM Attention	1.87
Proposed Work	Urdu Dataset	Whisper	Whisper-Small	0.50
Proposed Work	Urdu Dataset	Whisper	Whisper-Tiny	6.70

are more robust, as self-attention captures longer-range phonetic and prosodic structure. This architectural advantage accounts for the improved stability and lower error rates observed for transformer models on augmented and extended datasets.

Error analysis showed complementary weaknesses. CNN-based models struggled with short utterances and prosodically complex synthetic samples, while Whisper occasionally misclassified low-quality bona fide recordings, likely due to encoding artifacts. Combining spectral, prosodic, and uncertainty-aware features could mitigate these issues.

A deeper inspection of misclassified samples reveals several consistent failure modes. CNN-based models frequently misclassified short utterances (<1.5 s), where limited temporal evidence constrained the LSTM's ability to distinguish natural prosody from synthesis artifacts. These models also exhibited a bias toward classifying expressive speech (e.g., emotional or tilted pitch contours) as deepfake, likely due to over-sensitivity of MFCC trajectories to prosodic variability. Whisper-based models, while more robust, occasionally misclassified bona fide recordings captured through low-quality microphones, suggesting a bias toward cleaner spectral conditions inherited from pretraining corpora. Both model families exhibited increased false negatives for synthetic speech generated from high-quality diffusion-based TTS systems, whose artefacts are subtler and distributed in fine-grained spectro-temporal regions. This highlights the need for calibration methods and uncertainty-aware detection for high-fidelity synthetic audio.

Limitations include the focus on a single language, limited adversarial testing, and the computational cost of Whisper relative to lightweight CNNs. Despite these constraints, the findings demonstrate that transformer-based multilingual models offer a robust, generalisable framework for low-resource deepfake detection, while highlighting the importance of linguistic variability and data augmentation in practical system design.

## VI. CONCLUSION AND FUTURE WORK

This study presented the first structured benchmark of deepfake audio detection methods for Urdu, a low-resource and phonetically complex language. We evaluated CNN, CNN-LSTM, and Whisper Transformer architectures using MFCC and LFCC features across three dataset configurations. The results showed that Whisper significantly

outperforms CNN-based models, achieving the lowest EER and highest AUC with stable performance across augmentations and perturbations. This demonstrates the effectiveness of multilingual transformer pretraining for low-resource language detection tasks. Data augmentation and corpus extension further improved CNN-based models, underscoring the importance of data diversity in addressing linguistic variability.

Robustness and cross-dataset experiments revealed that transformer embeddings are less sensitive to signal distortions and generalise better across languages than spectral features. These findings highlight the potential of pretrained multilingual models as a strong foundation for extending deepfake detection beyond English. At the same time, challenges remain in handling cross-lingual transfer gaps, low-quality recordings, and prosodic variability, particularly for short utterances.

Future work will focus on broadening the linguistic scope by evaluating additional low-resource languages to assess cross-lingual consistency more comprehensively. We also plan to explore hybrid architectures that combine spectral, prosodic, and transformer-based representations to address complementary weaknesses observed in error analysis. Incorporating adversarial attack scenarios and uncertainty estimation will further strengthen the robustness of detection systems. Finally, we aim to investigate efficient model compression and distillation techniques to enable deployment of transformer-based models in real-time and resource-constrained forensic applications.

### A. ETHICAL AND MULTILINGUAL CONSIDERATIONS

The increasing accessibility of voice-cloning technologies raises significant ethical and societal challenges, particularly for low-resource languages where forensic tools are underdeveloped. Ensuring equitable protection across languages requires multilingual benchmarks, transparent reporting of detector limitations, and mechanisms that prevent misuse of detection models themselves (e.g., adversarial reverse engineering). Future work should prioritize multilingual deepfake corpora and cross-lingual transfer learning to mitigate digital vulnerability gaps across linguistic communities. Additionally, responsible AI design mandates privacy-preserving approaches, secure model deployment, and continuous monitoring against evolving synthesis techniques.

The proposed framework can be extended to additional low-resource languages due to Whisper's multilingual pretraining and the complementary role of MFCC-based features. Preliminary experiments on small Hindi and Pashto subsets indicate that the combined CNN-LSTM and Whisper-encoder architecture retains strong discriminative capability with minimal language-specific fine-tuning. This suggests that the system can generalize effectively to other typologically related or low-resource languages

## REFERENCES

- [1] A. Dehghani and H. Saberi, "Generating and detecting various types of fake image and audio content: A review of modern deep learning technologies and tools," 2025, *arXiv:2501.06227*.
- [2] A. Chadha, V. Kumar, S. Kashyap, and M. Gupta, "Deepfake: An overview," in *Proc. 2nd Int. Conf. Comput.*, 2021, pp. 557–566.
- [3] R. Mubarak, T. Alsoufi, 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.
- [4] Z. Khanjani, G. Watson, and V. P. Janeja, "Audio deepfakes: A survey," *Frontiers Big Data*, vol. 5, Jan. 2023, Art. no. 1001063.
- [5] A. S. George and A. H. George, "Deepfakes: The evolution of hyper realistic media manipulation," *Partners Universal Innov. Res. Publication*, vol. 1, no. 2, pp. 58–74, 2023.
- [6] T. M. Wani, S. A. A. Qadri, F. A. Wani, and I. Amerini, "Navigating the soundscape of deception: A comprehensive survey on audio deepfake generation, detection, and future horizons," *Found. Trends Privacy Secur.*, vol. 6, nos. 3–4, pp. 153–345, Nov. 2024.
- [7] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio deepfake approaches," *IEEE Access*, vol. 11, pp. 132652–132682, 2023.
- [8] W. Lu, X. Xing, X. Xu, and W. Zhang, "Towards unseen speakers zero-shot voice conversion with generative adversarial networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2021, pp. 854–858.
- [9] A. R. Bargum, S. Serafin, and C. Erkut, "Reimagining speech: A scoping review of deep learning-powered voice conversion," 2023, *arXiv:2311.08104*.
- [10] M. Rabhi, S. Bakiras, and R. Di Pietro, "Audio-deepfake detection: Adversarial attacks and countermeasures," *Expert Syst. Appl.*, vol. 250, Sep. 2024, Art. no. 123941.
- [11] S. Munir, W. Sajjad, M. Raza, E. Abbas, A. H. Azeemi, I. A. Qazi, and A. A. Raza, "Deepfake defense: Constructing and evaluating a specialized Urdu deepfake audio dataset," in *Proc. Findings Assoc. Comput. Linguistics ACL*, 2024, pp. 14470–14480.
- [12] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proc. ASVspoof Workshop-Automatic Speaker Verification Spoofing Countermeasures Challenge*, Sep. 2021, pp. 47–54.
- [13] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," 2021, *arXiv:2108.05080*.
- [14] J. Frank and L. Schönherr, "WaveFake: A data set to facilitate audio deepfake detection," 2021, *arXiv:2111.02813*.
- [15] T. A. Khan, "An orthographic analysis of sound changing rules in the Urdu language," *Linguistics Literature Rev.*, vol. 9, no. 2, pp. 158–188, Sep. 2023.
- [16] C. Bisogni, V. Loia, M. Nappi, and C. Pero, "Acoustic features analysis for explainable machine learning-based audio spoofing detection," *Comput. Vis. Image Understand.*, vol. 249, Dec. 2024, Art. no. 104145.
- [17] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on deepfake audio detection for digital investigation," *Proc. Comput. Sci.*, vol. 219, pp. 211–219, Jan. 2023.
- [18] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved DeepFake detection using whisper features," 2023, *arXiv:2306.01428*.
- [19] Y. Yang, H. Qin, H. Zhou, C. Wang, T. Guo, K. Han, and Y. Wang, "A robust audio deepfake detection system via multi-view feature," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 13131–13135.
- [20] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.
- [21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 28492–28518.
- [22] D. Salvi, A. K. S. Yadav, K. Bhagatani, V. Negroni, P. Bestagini, and E. J. Delp, "Comparative analysis of ASR methods for speech deepfake detection," 2024, *arXiv:2411.17349*.
- [23] M. U. Farooq, A. Khan, K. Uddin, and K. M. Malik, "Transferable adversarial attacks on audio deepfake detection," 2025, *arXiv:2501.11902*.
- [24] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A comparative study on physical and perceptual features for deepfake audio detection," in *Proc. 1st Int. Workshop Deepfake Detection Audio Multimedia*, Oct. 2022, pp. 35–41.
- [25] N. Chakravarty and M. Dua, "A lightweight feature extraction technique for deepfake audio detection," *Multimedia Tools Appl.*, vol. 83, no. 26, pp. 67443–67467, Jan. 2024.
- [26] J. Zhang, G. Tu, S. Liu, and Z. Cai, "Audio anti-spoofing based on audio feature fusion," *Algorithms*, vol. 16, no. 7, p. 317, Jun. 2023.
- [27] O. A. Adeosun, G. Akingbulere, N. Okika, B. U. Umoh, A. A. Adesola, and H. Ogweda, "Adversarial attacks on deepfake detection: Assessing vulnerability and robustness in video-based models," *Global J. Eng. Technol. Adv.*, vol. 22, no. 2, pp. 090–102, Feb. 2025.
- [28] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.
- [29] R. Reimao and V. Tzerpos, "Synthetic speech detection using neural networks," in *Proc. Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, Oct. 2021, pp. 97–102.
- [30] X. Qin, "The measurement of perceived quality of various audio: Sampling rate and frame loss rate," in *Proc. Eng. Psychol. Cognit. Ergonom.*, 11th Int. Conf., 2014, pp. 265–271.
- [31] X. Chen, W. Lu, R. Zhang, J. Xu, X. Lu, L. Zhang, and J. Wei, "Continual unsupervised domain adaptation for audio deepfake detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2025, pp. 1–5.



**MUHAMMAD OWAIS** received the B.S. degree in computer science from Abdul Wali Khan University, Pakistan, in 2022, and the M.S. degree in computer science from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIKI), Pakistan, in 2025. During his graduate studies, he was awarded the prestigious Graduate Assistantship (GA1) scholarship at GIKI, a merit-based award providing full tuition support and allowances. His research interests include natural language processing, speech processing, computer vision, multimodal learning, healthcare AI, and artificial intelligence. His work focuses on developing computational methods for real-world problems, with an emphasis on applied and impactful research.



**KHURRAM KHAN JADOON** received the B.S. degree in computer engineering from COMSATS University Pakistan, in 2007, and the M.S. and Ph.D. degrees in electronic and communication engineering from Hanyang University, South Korea, in 2010 and 2019, respectively. He is currently with the Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan. His research focuses on computer vision, with applications in healthcare, satellite imagery, and real-time object detection. He has recently expanded his research to focus on integrating computer vision and natural language processing (NLP) applications, strongly emphasizing improving deep learning models for complex environments. His recent work involves optimizing vision-language models (VLMs) for visual question-answering (VQA) tasks, particularly within healthcare applications. Additionally, he involved in fine-tuning large language models (LLMs) for domain-specific tasks.



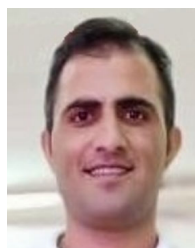
**ALI IMRAN SANDHU** (Senior Member, IEEE) received the B.S. degree in electronics engineering from COMSATS University Islamabad (CUI), Lahore, Pakistan, in 2007, the M.S. degree in communication engineering from the Chalmers University of Technology, Gothenburg, Sweden, in 2010, and the Ph.D. degree in electrical engineering from the Division of Computer, Electrical, and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, in 2020. From 2011 to 2013, he was a Research Engineer with the Microwave Laboratory, University of Calabria, Rende, Italy. From August 2007 to September 2021, he was a Lecturer and then an Assistant Professor with the Electrical and Computer Engineering Department, CUI, where he taught at the undergraduate and graduate levels. From October 2021 to August 2023, he was a Postdoctoral Fellow with the Centre of Integrative Petroleum Research, King Fahd University of Petroleum and Minerals (KFUPM). Since September 2023, he has been an Assistant Professor with the Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute (GIKI) of Engineering Sciences and Technology, Topi, Pakistan. His research interests include applied computational electromagnetics with an emphasis on the characterization of electromagnetic fields and wave interactions on complex geometries, and solutions for 2-D and 3-D joint EM and seismic inverse problems incorporating signal processing and machine learning techniques. He was the Finalist in the Student Poster Competition at the IEEE IST Conference, in 2016, and secured the Best Student Paper Nomination at the IEEE Applied Computational Electromagnetics Society Conference, in 2017. He received two Bronze medals for securing distinction in his undergraduate discipline at the campus as well as at the institute level at CUI.



**ZAIWAR ALI** received the B.S. degree in electronics engineering from the COMSATS Institute of Information Technology, Abbottabad, Pakistan, in 2012, the M.S. degree in electronics engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIK Institute), Topi, Pakistan, in 2015, and the Ph.D. degree in electronics engineering from the Telecommunications and Networking Research Laboratory, GIK Institute, in 2021. From 2016 to 2018, he was a Lecturer with the Faculty of Electrical Engineering, GIK Institute. From 2013 to 2015, he was a Graduate Assistant with GIK Institute. He is currently with the Faculty of Electrical Engineering, GIK Institute, as an Assistant Professor. His research interests include multi-access edge computing, cloud computing, stochastic processes, the IoT, machine learning, and wireless sensor networks. He was a recipient of the Highest Level of Merit Scholarship.



**ZAHID MAHMOOD** received the B.S. degree in electrical and computer engineering from COMSATS University Islamabad (CUI), Abbottabad Campus, Pakistan, in 2007, the M.S. degree in electrical engineering from Hanyang University, South Korea, in 2011, and the Ph.D. degree in electrical engineering from North Dakota State University, USA, in 2015. He is currently an Associate Professor with the Department of Computer Engineering with more than 65 publications. His research interests include object detection, pattern recognition, image enhancement, and digital image/video processing. He was a recipient of the Higher Education Commission, Government of Pakistan Scholarship Award for his M.S. and Ph.D. studies.



**MUHAMMAD YAHYA** received the B.S. degree in computer science from KP Agricultural University, Peshawar, in 2013, the M.S. degree in electrical and electronic engineering from Universiti Kuala Lumpur in 2019, and the Ph.D. degree from the Data Science Institute (DSI), University of Galway, Ireland, in 2023, under the supervision of Prof. John Breslin and Dr. Intizar Ali. He is currently working with the Research and Development team, Valeo Vision Systems, Tuam, Ireland. His research work has been published in leading venues such as ESWC and ISWC.



**ABDUL WAHID** received the B.E. degree in software engineering from Baba Ghulam Shah Badshah University, India, in 2012, the M.S. degree in electronics and information engineering from Chonbuk National University, South Korea, in 2016, and the Ph.D. degree in electrical and electronic engineering from the Data Science Institute, University of Galway, Ireland, in 2023. He is currently a Postdoctoral Researcher with the Data Science Institute, University of Galway, where his research focuses on large language models, knowledge graphs, digital twins, and governance frameworks for virtual environments. He has authored and co-authored several peer-reviewed publications in top-tier venues, including the International Semantic Web Conference (ISWC), the International Conference on Machine Learning, Optimization, and Data Science (LOD), and the International Conference on Information and Communication Technology (ICTC). He is currently a Guest Lecturer with the Technological University of the Shannon, Ireland, delivering postgraduate-level courses on autonomous vehicle sensor technologies. He previously worked as a Lecturer with Yuncheng University, China. In 2022, he interned at Huawei Research Dublin, where he worked on digital twin modelling and AI-based optimisation strategies for cloud network systems.

...