Forgery Detection Approaches for AI-Generated Speech

1️⃣ Wav2Vec2-Based Fine-Tuning

Key Innovation:

- Uses self-supervised learning to extract speech representations from raw audio.
- Fine-tuned on deepfake audio datasets to classify real vs. synthetic voices.

Performance Metrics:

- Accuracy: ~90% (varies based on dataset)
- High recall in detecting synthetic speech.

Why This is Promising:

✅ Pre-trained on large-scale speech data, reducing labeled data requirements.
✅ Capable of processing real-time audio efficiently.
✅ Robust against low-quality or compressed audio.

Potential Challenges:

⚠️ Requires domain adaptation for specific deepfake synthesis techniques.
⚠️ Computationally expensive for real-time inference.

2️⃣ Spectrogram-Based CNN (Convolutional Neural Networks)

Key Innovation:

- Converts audio into spectrogram images and applies CNNs for classification.
- Detects subtle artifacts in frequency patterns introduced by AI-generated speech.

Performance Metrics:

- F1-Score: ~92%
- Works well with **limited training data.

Why This is Promising:

✅ CNNs excel at pattern recognition, capturing deepfake artifacts.
✅ Can be optimized for real-time processing.
✅ Works across different languages & speakers.

Potential Challenges:

⚠️ May struggle with **highly compressed or noisy audio**.
⚠️ CNNs need **careful tuning** for new deepfake generation techniques.
3️⃣MFCC + LSTM (Mel-Frequency Cepstral Coefficients + Long Short-Term Memory)

Key Innovation:

- Extracts MFCC features (mimicking human auditory perception).
- Uses LSTM to model speech temporal dependencies, detecting unnatural transitions.

Performance Metrics:

- Accuracy: 87-93% (depending on dataset).
- Performs well on short-duration audio clips.

Why This is Promising:

✅ Lightweight model, suitable for real-time detection.
✅ Captures long-term dependencies, making it robust for conversational deepfakes.
✅ Adaptable to different deepfake architectures.

Potential Challenges:

⚠️ Sensitive to background noise and requires preprocessing.
⚠️ Less effective for very short speech clips (<1 sec).

Final Thoughts:

- Wav2Vec2 is great for generalization and high accuracy.
- CNN-based approaches are excellent for catching spectral anomalies.
- MFCC + LSTM is best for lightweight real-time applications.

A hybrid approach combining multiple techniques might yield the best results. 🚀