Part 1: Research & Selection

1. Real-Time ResNet/LCNN System

(Source: arXiv study on communication platforms[1])

Key Innovation:

- Dual architecture using ResNet (for LA spoofing) and LCNN (for PA spoofing)
- Mel-spectrogram/power-spectrogram feature extraction optimized for streaming audio
- Executable cross-platform software for real-time deployment

Performance:

- Achieved EER (Equal Error Rate) of 0.83% on ASVspoof 2019 LA dataset
- Tested in live Microsoft Teams meetings with <200ms latency

Why Promising:

☑ Designed specifically for continuous conversational speech analysis

☑ Verified in actual communication platform environments

☑ Modular architecture allows integration with existing voice pipelines

Limitations:

⚠ Requires audio trimming/padding to fixed 6-second windows

⚠ Performance drops with background noise >45dB SNR

2. VGG16-LSTM Hybrid Model

(Source: IJCRT paper[2])

Key Innovation:

- Combines VGG16's spatial feature extraction with LSTM temporal analysis
- Augments MFCCs with handcrafted features (spectral centroid, roll-off)
- Ensemble learning with XGBoost classifier

Performance:

- 98.2% accuracy on ASVspoof 2019
- 0.91 F1-score for real-time classification

Why Promising:

☑ Processes raw audio streams without pre-segmentation

☑ Feature fusion captures both vocal tract and prosody characteristics

☑ Lightweight enough for edge deployment (2.1M parameters)

Limitations:

⚠ Requires GPU acceleration for real-time performance

⚠ Vulnerable to adversarial attacks using phase reconstruction

3. Contrastive Learning Detector (CLAD)

(Source: arXiv robust detection study[3])

Key Innovation:

- Contrastive learning framework resistant to 23 audio manipulation attacks
- Length loss regularization for variable-duration inputs
- Frequency-domain adversarial training

Performance:

- 98.7% accuracy on manipulated ASVspoof samples
- FAR <1.63% against volume/fading/reverb attacks

Why Promising:

☑ Specifically hardened against evasion techniques

☑ No preprocessing needed - works on raw waveforms

☑ Maintains 94% accuracy in noisy environments (15dB SNR)

Limitations:

⚠ Requires retraining for new attack vectors

⚠ Higher computational load than traditional CNNs

Selection Rationale

These approaches were chosen for their:

1. **Real-Time Capability** - All demonstrate sub-second inference times
2. **Robustness** - Complementary defenses against different attack types
3. **Deployability** - Include implementation frameworks beyond pure accuracy metrics

For implementation, I recommend starting with the VGG16-LSTM hybrid[2] as it balances accuracy with moderate computational requirements, while planning integration of CLAD's contrastive learning[3] for adversarial robustness. The real-time ResNet/LCNN system[1] provides immediate deployable architecture reference.

1. https://arxiv.org/html/2403.11778v1
2. https://www.ijcrt.org/papers/IJCRT24A4745.pdf
3. https://arxiv.org/html/2404.15854v1