Audio Classification - Model layer

Hugging Face model zoo options (good starting checkpoints)

Transformers on spectrograms (patch-based ViTs)

- **AST (Audio Spectrogram Transformer)** — `MIT/ast-finetuned-audioset-10-10-0.4593` *Multi-label Audioset head; strong general audio tagging baseline.*
- **PaSST (Patchout Spectrogram Transformer)** — e.g., `kkoutini/passt_s_kd_ast10_10` *Patchout regularization → efficient + robust on long clips.*
- **HTS-AT (Token-Semantic Transformer for Audio Tagging)** — `TencentGameMate/chinese-htsat` (family) *Strong music/environmental tagging; pairs well with CLAP text heads.*
- **Audio-MAE (Masked Autoencoding pretrain on spectrograms)** — e.g., `MIT/ssast-base-patch400` (SSAST/AudioMAE family) *Self-supervised pretraining → label-efficient fine-tuning.*

Self-supervised waveform encoders (SSL)

- **Wav2Vec2 / HuBERT / WavLM** — `facebook/wav2vec2-base`, `superb/hubert-large-superb-er`, `microsoft/wavlm-base-plus` *Encode raw wave; add small classifier head for KWS/ASC/emotion.*
- **BEATs** — `microsoft/BEATs` / `microsoft/BEATs-iter3` *SSL optimized for non-speech acoustic events; strong ESC-50/ASC results.*

CNN baselines (fast, small, deployable)

- **PANNs (CNN14/ResNet22)** — `qiuqiangkong/panns_cnn14` *Mature CNN family; good speed/quality trade-off.*
- **YAMNet (MobileNet-V1)** — ports on HF (search "yamnet") *Tiny, mobile-friendly; handy for real-time tagging.*

Audio–text contrastive (zero-shot & tagging)

- **CLAP (Contrastive Language–Audio Pretraining)** — `laion/clap-htsat-unfused` / `laion/clap-htsat-fused` *Zero-shot label search via text prompts; also fine-tunes to tags.*
- **AudioCLIP** — community ports (search "audioclip") *Image/audio/text shared space; useful for multimodal label spaces.*

Task-specific heads

- **Keyword spotting (Speech Commands)** — `superb/wav2vec2-base-superb-ks`
- **Acoustic scenes** — `dcase2020/task1a-baseline-*` (various community ports)

- **Music tagging** — `M-A-P/MERT-v1-95M` (music spectral transformer), `mtg-jamendo-*` adapters

> Pick by use-case: **Audioset-style multi-label** → AST/PaSST/HTS-AT/BEATs; **KWS** → Wav2Vec2/WavLM small; **On-device** → YAMNet/PANNs; **Zero-shot** → CLAP.

Architectural innovations (what actually moves the needle)

- **Spectrogram ViTs (AST/PaSST/HTS-AT):** Convert log-mels to **patch tokens**, use **self-attention** to model long contexts; **patchout** (PaSST) randomly drops patches for regularization & speed. Attention **pooling** over time/freq replaces fixed global pooling for better temporal localization.
- **Self-supervised waveform encoders (Wav2Vec2/HuBERT/WavLM/BEATs):** Pretrain on raw audio with contrastive/masked objectives → **strong universal features**; downstream = a small **classification head**. **BEATs** tailors SSL to **non-speech acoustic** semantics.
- **Weak-label learning & MIL pooling (AudioSet/FSD50K):** Use **segment-level tokens** with **MIL/attention pooling** to aggregate clip-level predictions; handles noisy/weak labels and long clips.
- **Audio–text contrastive (CLAP/AudioCLIP):** Align audio and natural-language labels → **zero-shot** classification; lets you "prompt" new classes without retraining.
- **Learnable front-ends (LEAF, SincNet):** Replace fixed mel filters with **learnable filterbanks**, improving robustness across domains/devices.
- **Multi-scale & efficient designs: Dilated CNNs**, **temporal pooling pyramids**, and **patch subsampling** (PaSST) keep **latency + memory** low for deployment.
