Classifying the Acoustic Environment of YouTube Videos using Deep Neural Networks

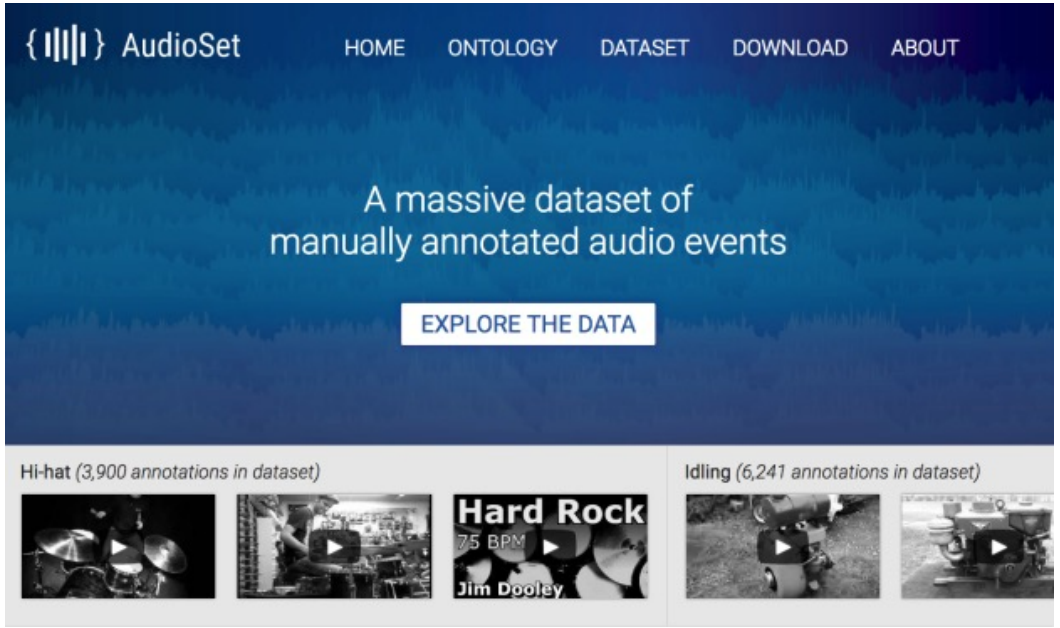
Andrew Chang, PhD

Unique Selling Proposition

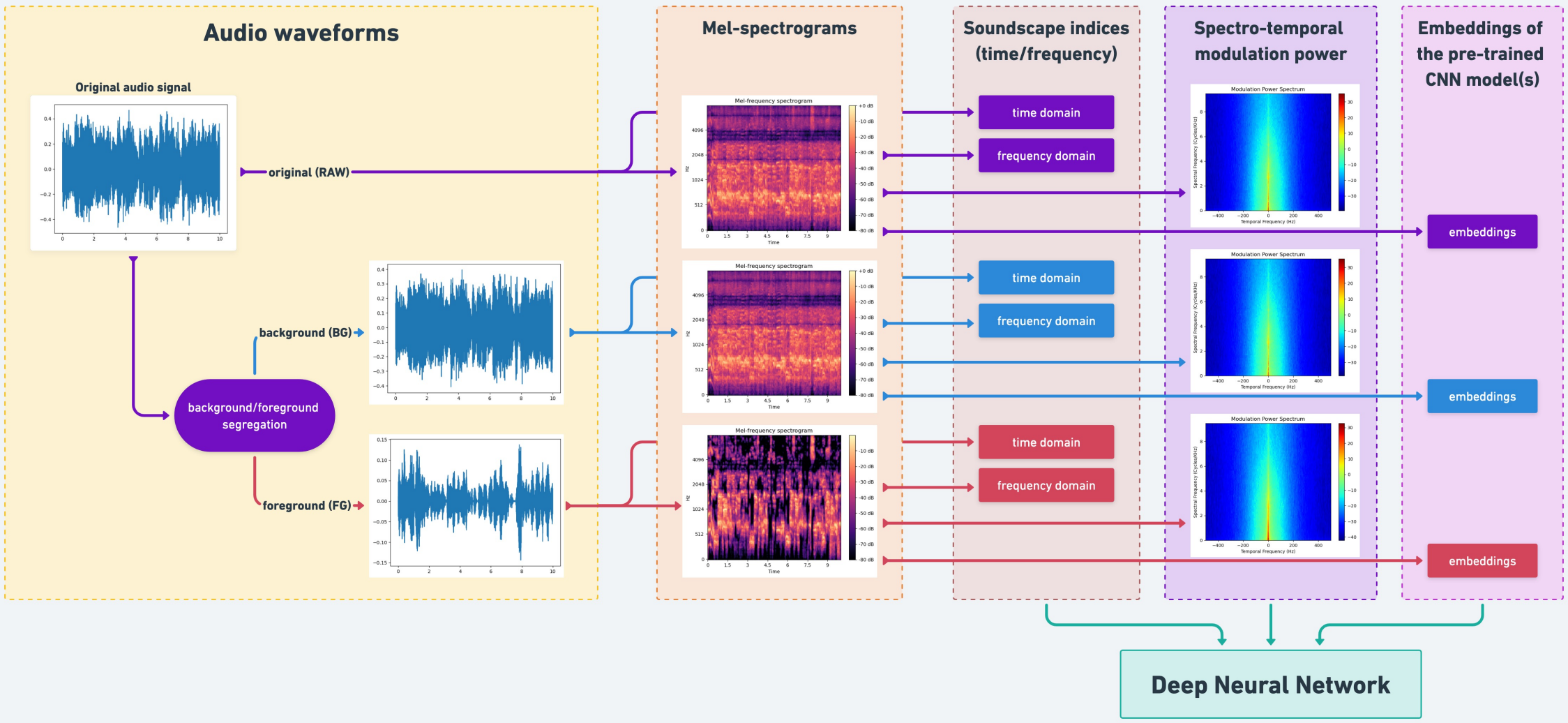
- Goal: classify the environment in which the audio was recorded as either urban or natural.
- Combined traditional audio signal processing and pre-trained audio CNN, achieved 90.1% adjusted accuracy (27.6% improvement) for classifying audio recording environment.

Google AudioSet

- Human-annotated 10-second sound clips sourced from YouTube videos
  - Natural environments : 6,862 recordings
  - Urban environments: 6,564 recordings
- Various recording settings and editing methods used in YouTube audios

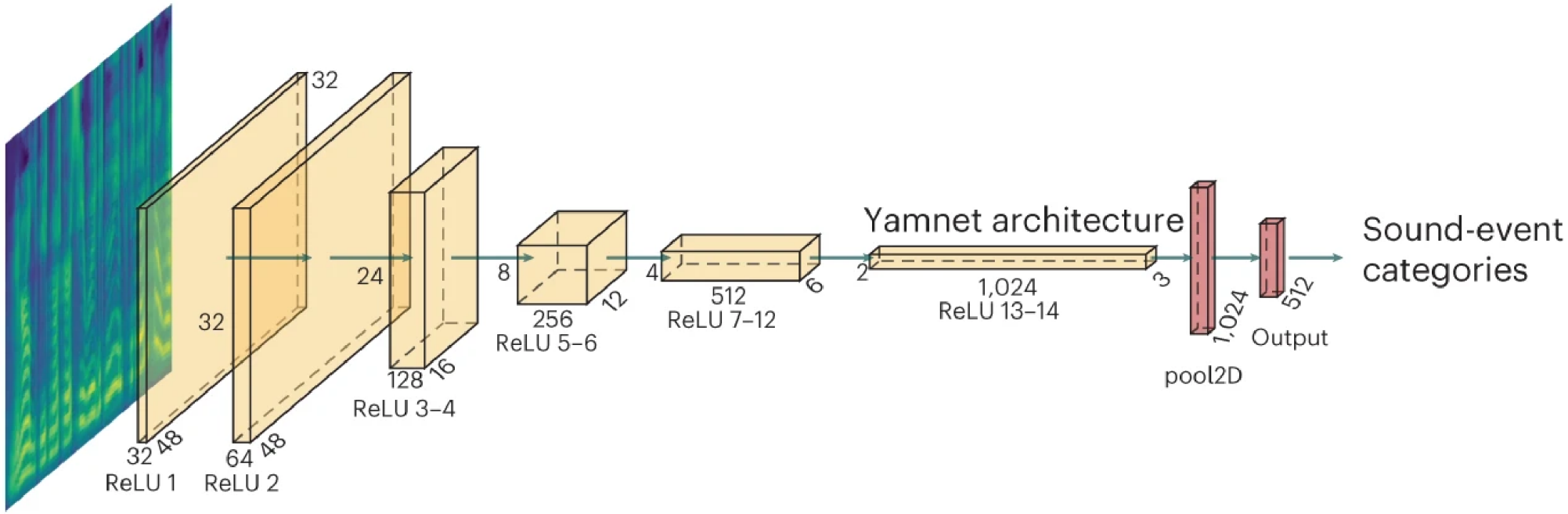


Data preprocessing

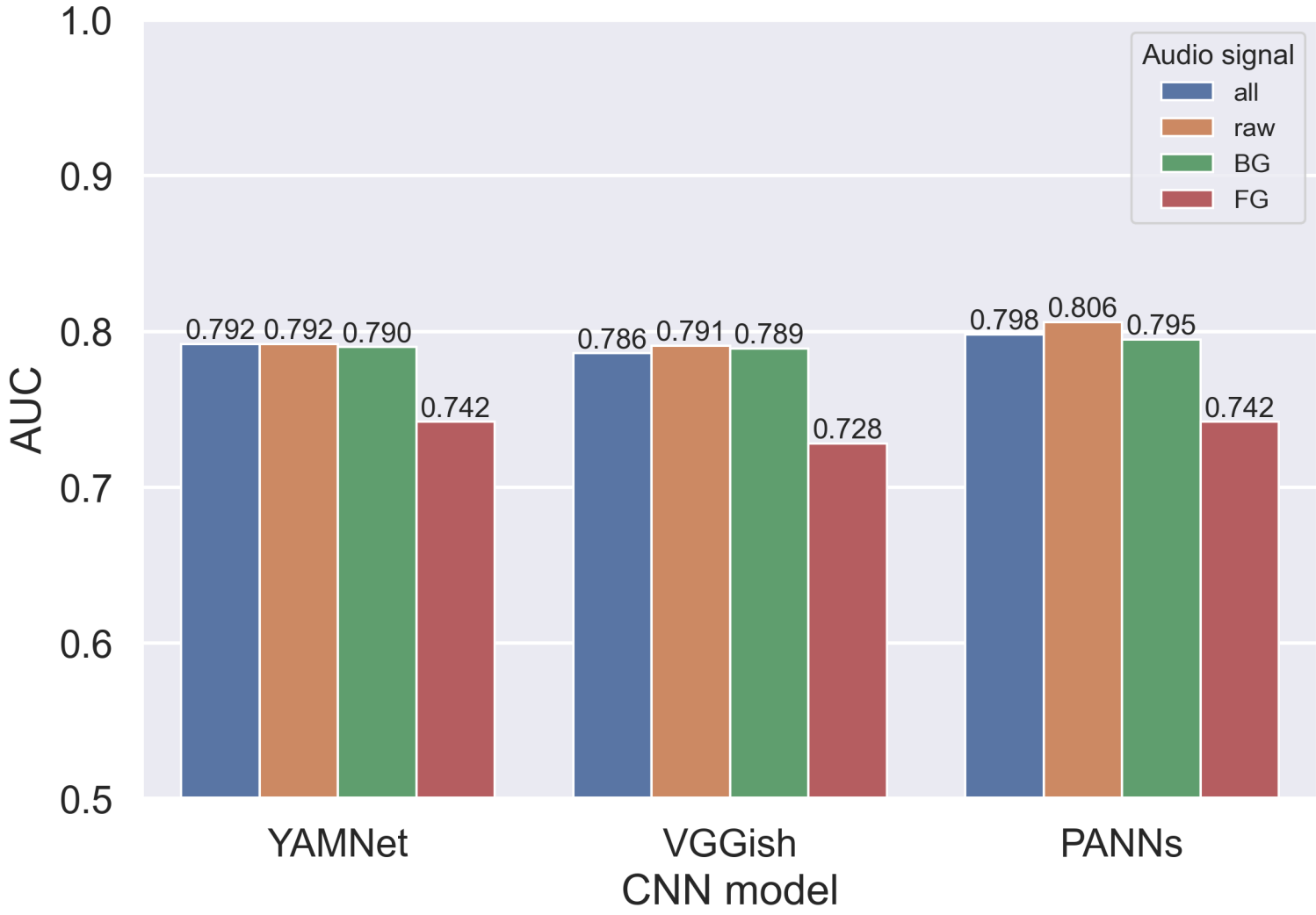


Convert Audio Signals to Pre-trained CNN Embeddings

- VGGish (256 embs), YAMNet (1024 embs), PANNs (2048 embs)
- Dimensional reduction using PCA

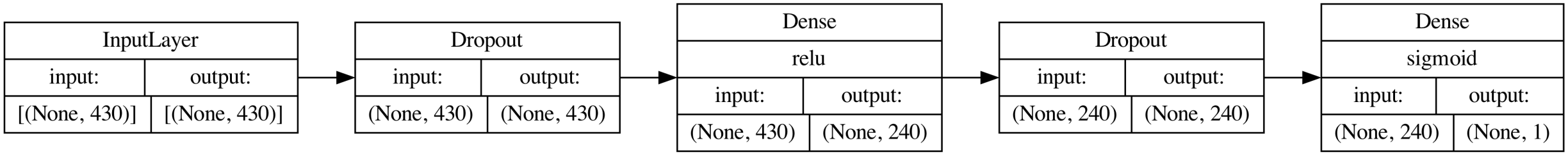


Bayesian hyperparameter tuning & feature selection



Best features: raw audio signals + PANNs embeddings

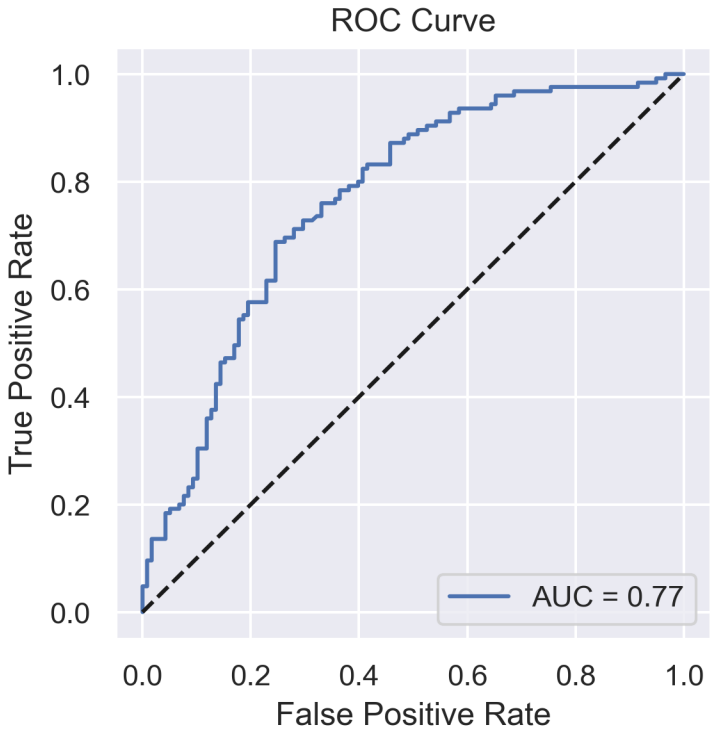
Hyperparameter-tuned multilayer perceptron model

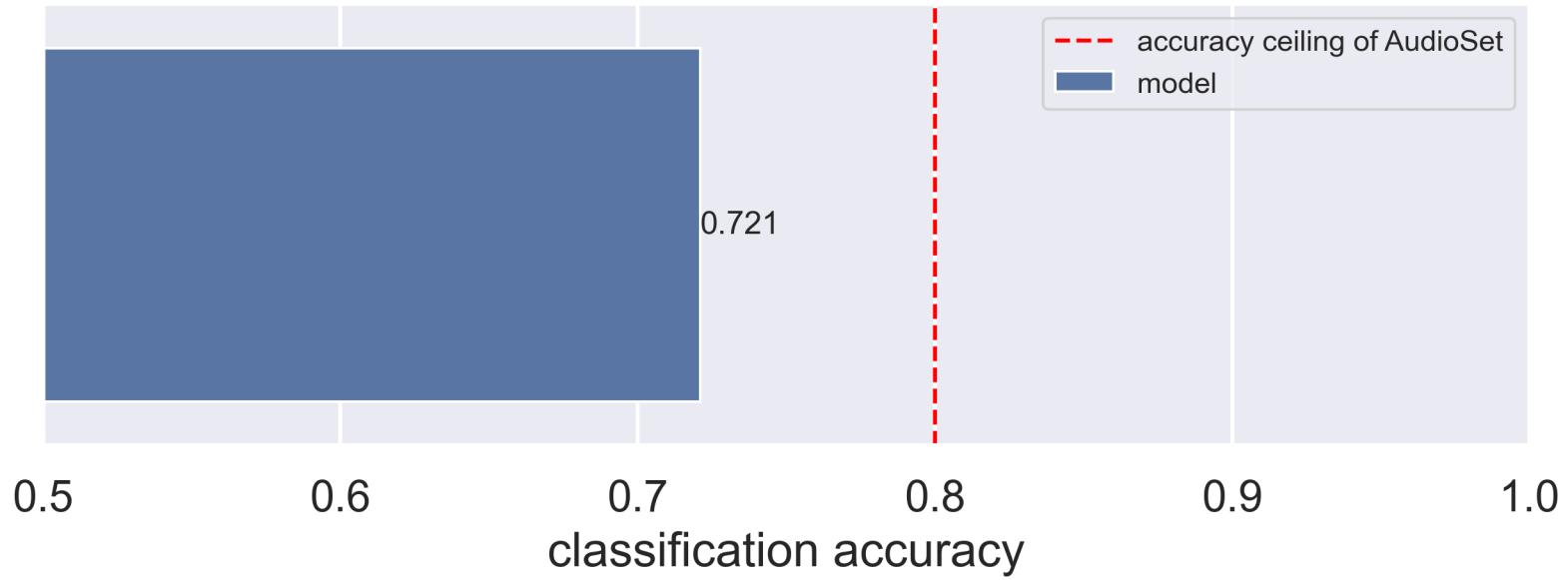


Performances on the testing dataset

ROC-AUC: 0.77

adjusted accuracy: 90.1%



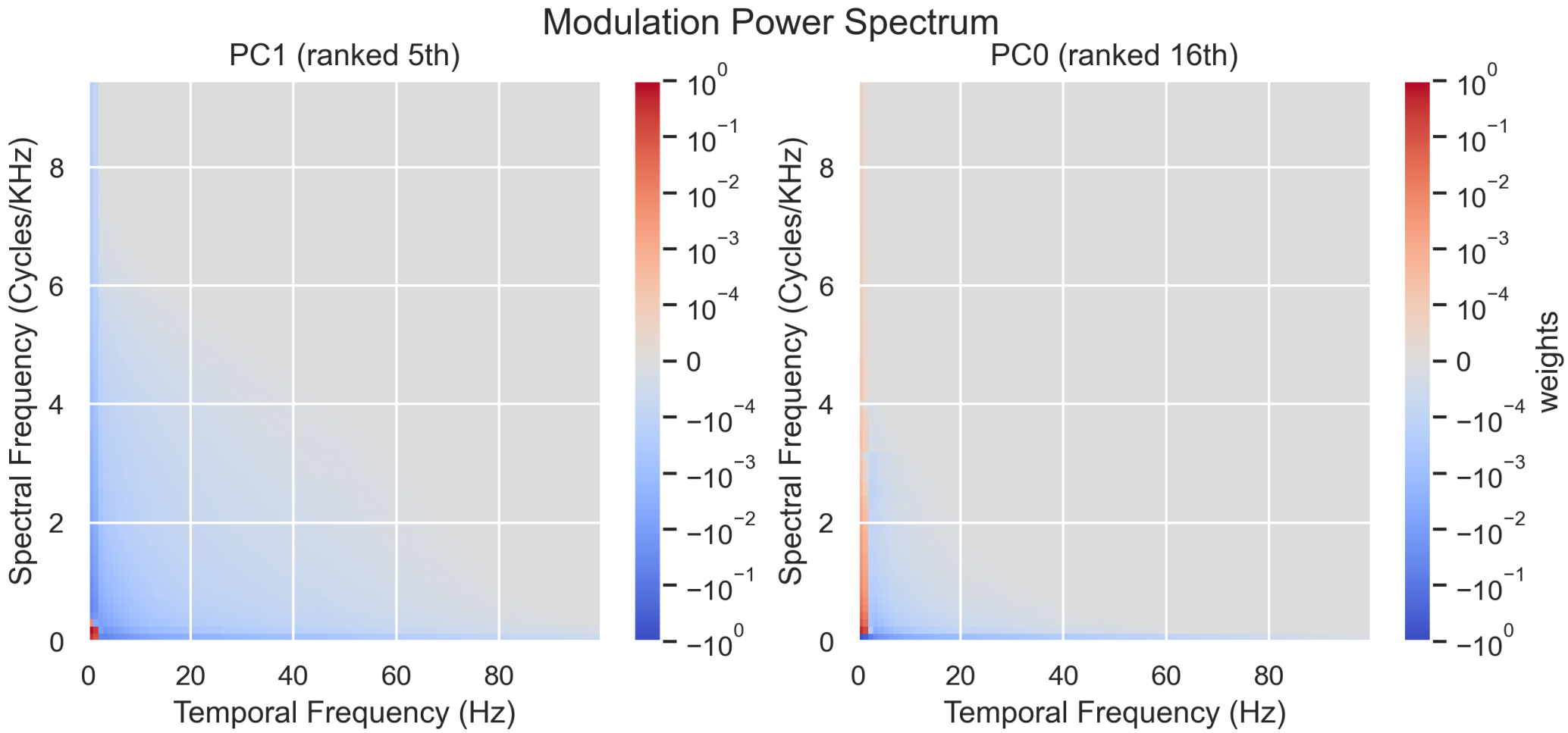


Feature importance

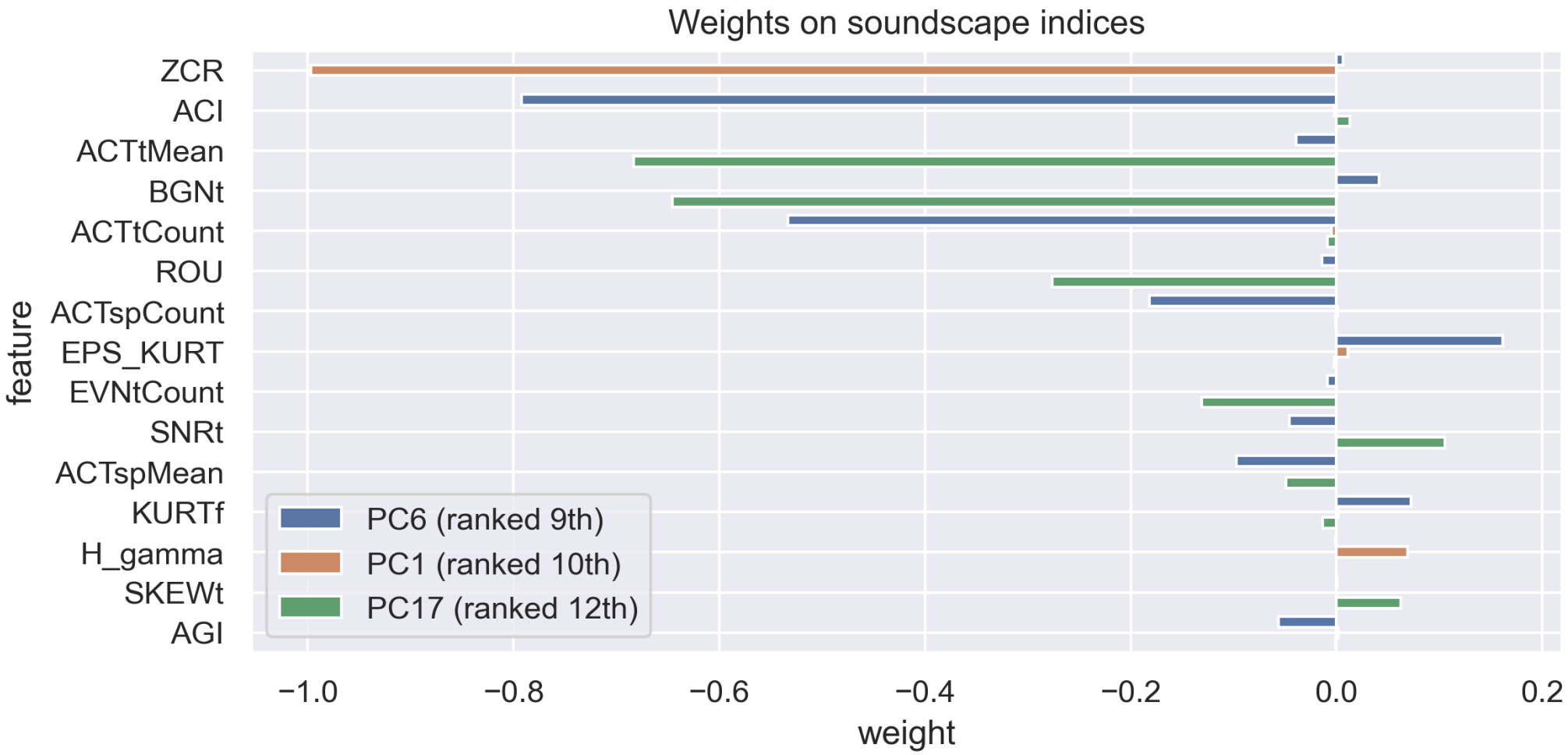
Among the top 20 PCs:

15 PCs of PANNs embeddings
3 PCs of soundscape indices
2 PCs of modulation power spectrum (MPS)

Important MPS features:
Low-frequency spectrotemporal modulations



Important soundscape indices



For more in-depth explanations and references of soundscape indices, please visit scikit-maad.

Conclusions

- A deep neural network model was created to classify YouTube videos as recorded in urban or natural environments.
- The model combined features from audio signal processing and pre-trained audio CNN models, achieving an AUC of 0.766.
- Pre-trained audio CNN model embeddings dominated the model's output, but traditional signal processing features were also informative.
- The model had low overfitting (0.035) and high adjusted accuracy (0.901).
- The zero-crossing rate, acoustic complexity index, and low-frequency spectrotemporal modulations were highly informative features for classifying urban or natural sounds.

Future directions

- Improving the dataset with more accurate and diverse environmental labels can further refine the model, making it more practical for classifying different acoustic environments, such as small rooms or cars.
- Combining the model's features with other audio recognition models and feeding them into a generative large language model, such as GPT, can lead to a semantic summary of the audio recording.