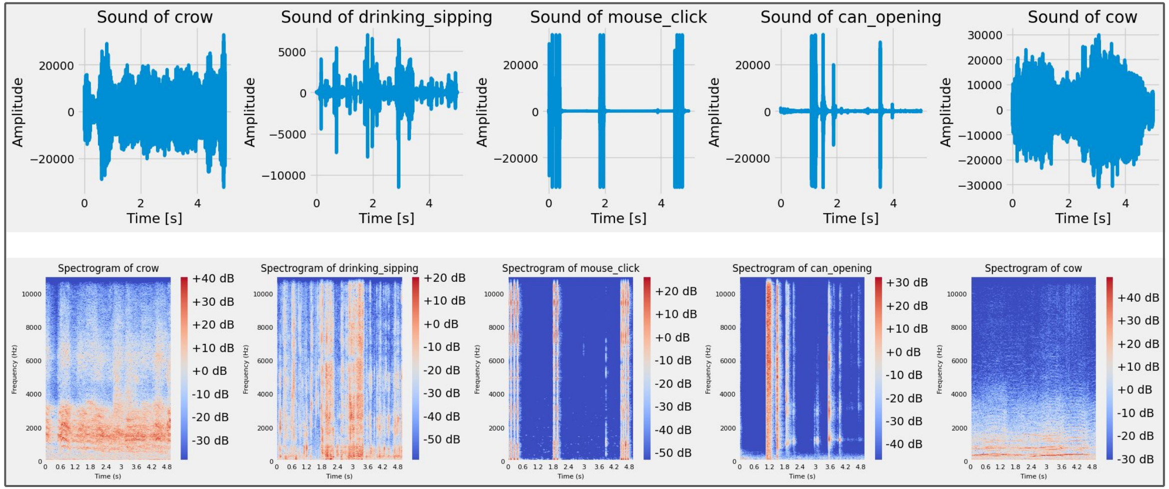
Environment Sound Classification

Saksham Singh Kushwaha (sxk230060)

Prof. Rishabh Iyer
(CS 6375)

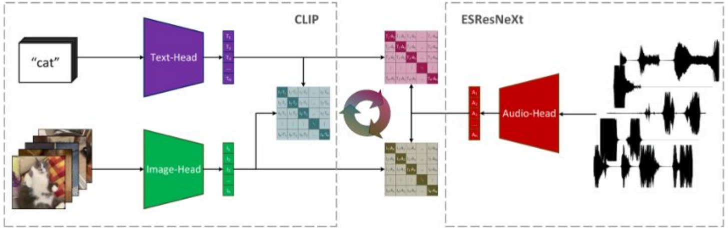
ESC-50 Data

- Standard dataset for environmental sound classification
- **2000** labeled audio recordings of **5-sec** length
- **50 classes** (40 per class); For eg. Animal, Human, interior sounds etc.
- Pre-arranged **5-folds** for cross-validation
- Few examples:

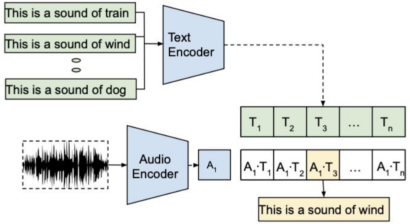


Experiments

- Compare several machine learning models
  - Supervised learning
    - Classical ML models
      - Logistic regression, Decision tree etc.
      - Low level features: frequency, pitch, loudness etc.
    - Deep learning models:
      - CNN on raw-audio and log-mel spectrograms
  - Zero-shot approach
    - AudioCLIP (Guzhov et. al)



Credits Guzhov et al. AudioCLIP: Extending CLIP to Image, Text and Audio



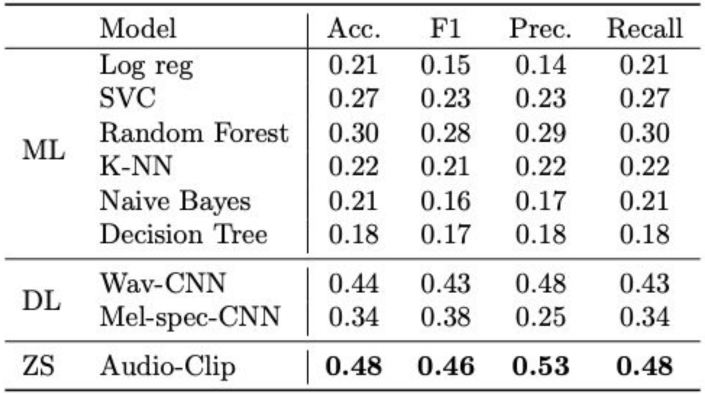
Zero-shot approach

Results

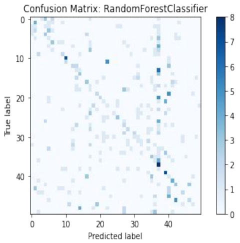
- Implementation:
  - ML: default parameters in sklearn
  - CNN: 50 epochs (pytorch)

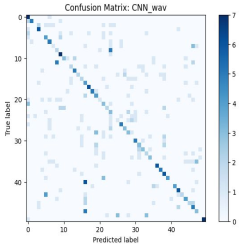
- Avg. of 5 fold cross-validation
- Performance: **ZS > DL > ML**
- Benefit of pretraining on large data
- Confusion matrix

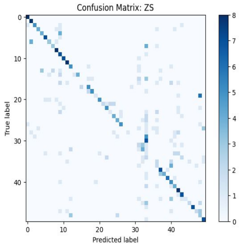
- Limitations and Future Work:
  - Hyper-param search with val-split
  - Data augmentation



	Model	Acc.	F1	Prec.	Recall
ML	Log reg	0.21	0.15	0.14	0.21
	SVC	0.27	0.23	0.23	0.27
	Random Forest	0.30	0.28	0.29	0.30
	K-NN	0.22	0.21	0.22	0.22
	Naive Bayes	0.21	0.16	0.17	0.21
	Decision Tree	0.18	0.17	0.18	0.18
DL	Wav-CNN	0.44	0.43	0.48	0.43
	Mel-spec-CNN	0.34	0.38	0.25	0.34
ZS	Audio-Clip	**0.48**	**0.46**	**0.53**	**0.48**


Confusion Matrix: RandomForestClassifier
True label
Predicted label


Confusion Matrix: CNN_wav
True label
Predicted label


Confusion Matrix: ZS
True label
Predicted label