Human Data Analytics 2023-24

Exploring CNN, LSTM, and Attention Mechanism with Augmentation Techniques on ESC-50 Dataset

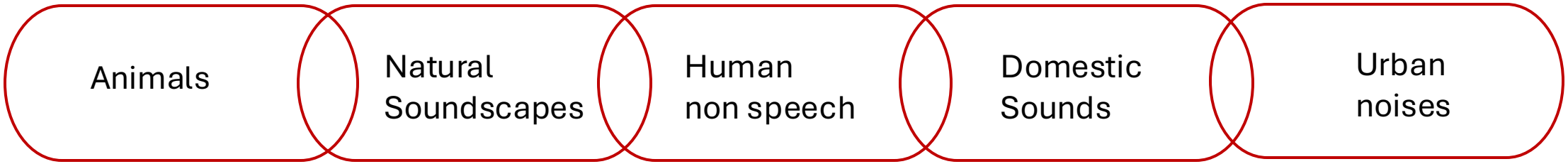
Roben Bhatti - 2091187

Goal

- **Classification task** of Environmental Sound
- Explore **deep learning** models using CNN, LSTM, Attention mechanism
- Use Data **Augmentations** techniques to Improve Accuracy and Dataset size

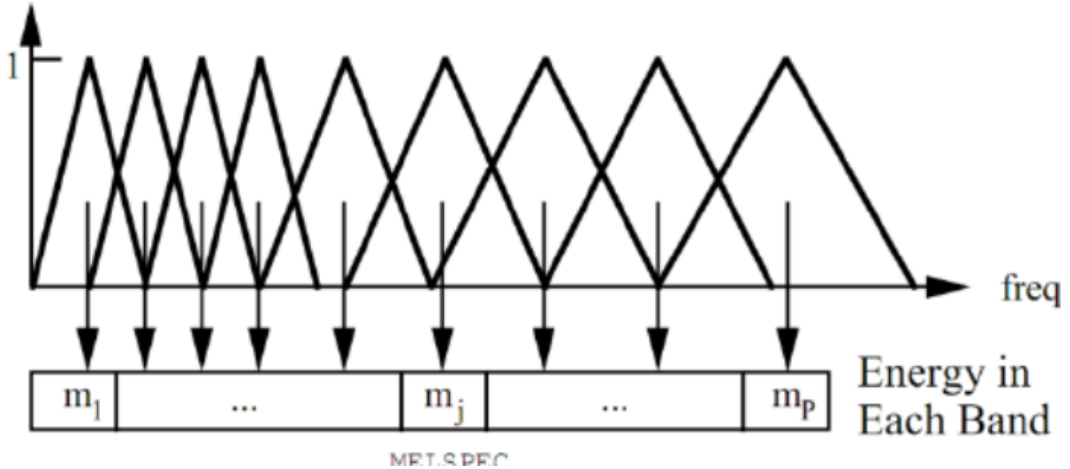
Esc-50 Dataset

- Collection of **2,000** labeled environmental audio recordings
- **5 seconds** long and categorized into **50 distinct semantic classes**
- Only 40 samples per class
- 5 Macro Categories

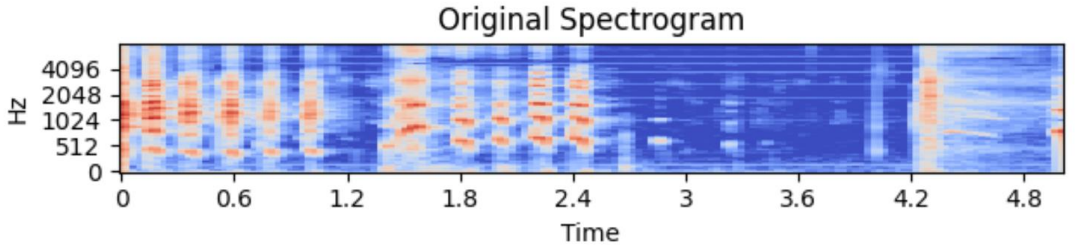


Mel Spectrogram

- **STFT** on overlapping windows
- **Hann** window to reduce spectral leakage
- **Mel filterbank** to align with human auditory perception



$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n] * w[n-m] * e^{-j\omega n}$$



Audio Augmentations

- **Gaussian Noise**: Randomly adds noise with a minimum amplitude of 0.001 and a maximum amplitude of 0.015.
- **Time Stretching**: Randomly speeds up or slows down the audio by a factor ranging from 0.8 to 1.25.
- **Pitch Shifting**: Shifts the pitch of the audio by up to 4 semitones in either direction.
- **Shifting**: Applies random temporal shifts to the audio with a maximum shift of 0.5 seconds.

- Probability of 50% for each filter

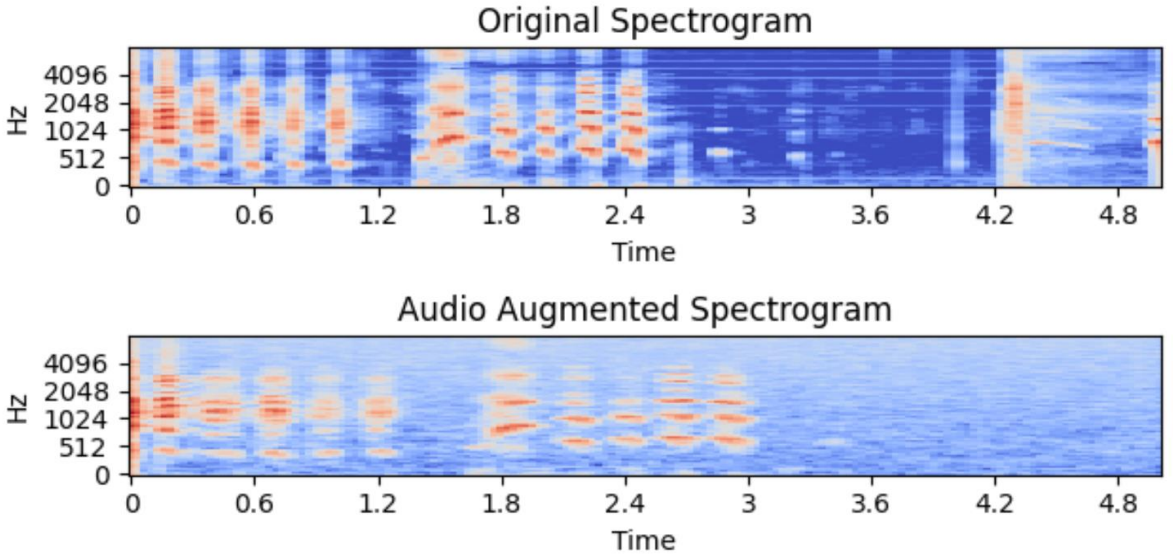
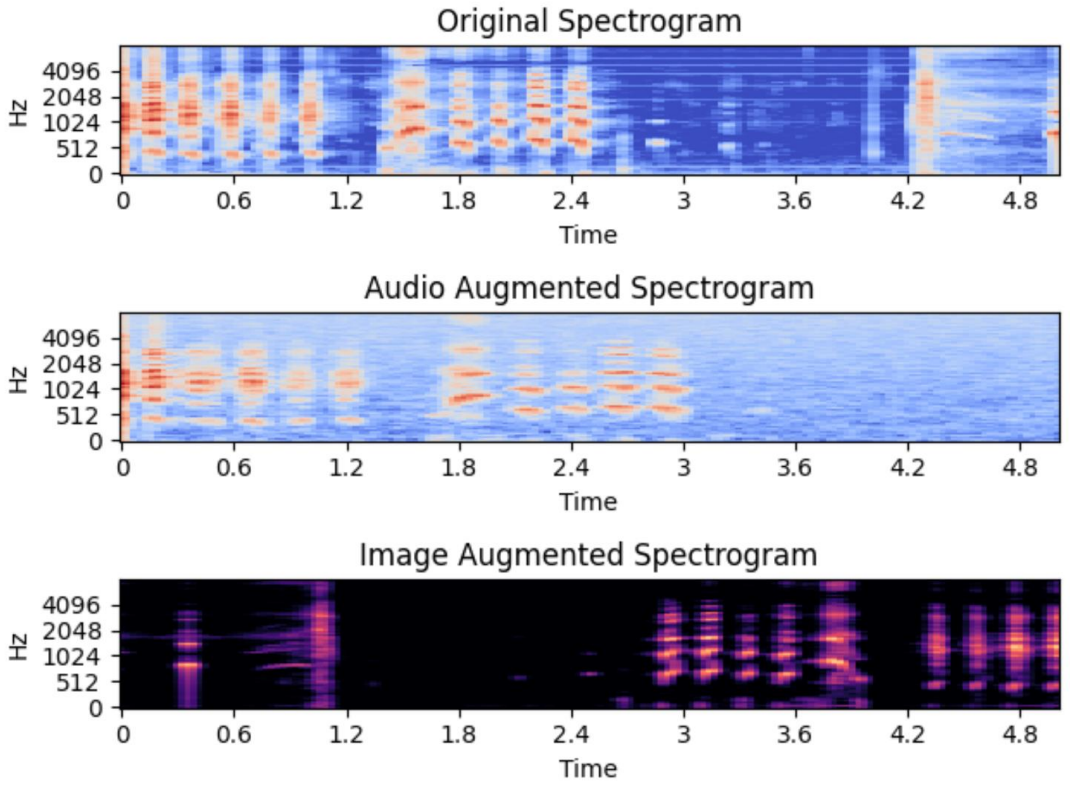


Image Augmentations

- **Random Horizontal Flip**: Randomly flips the spectrogram along the time axis.
- **Random Brightness Adjustment**: brightness ± 10%.
- **Random Translation**: Translates image by up to 10% in both the time and frequency axes.
- **Random Contrast Adjustment**: Alters the contrast by a factor between 0.6 and 1.4.

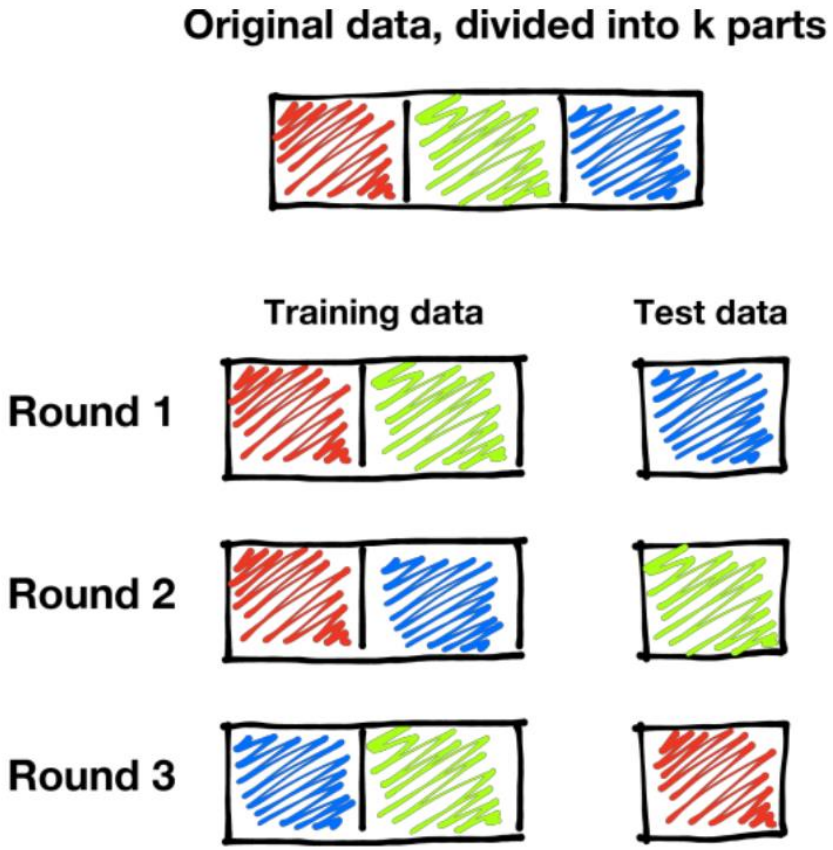


- Probability of 50% for each layer

Dataset splitting

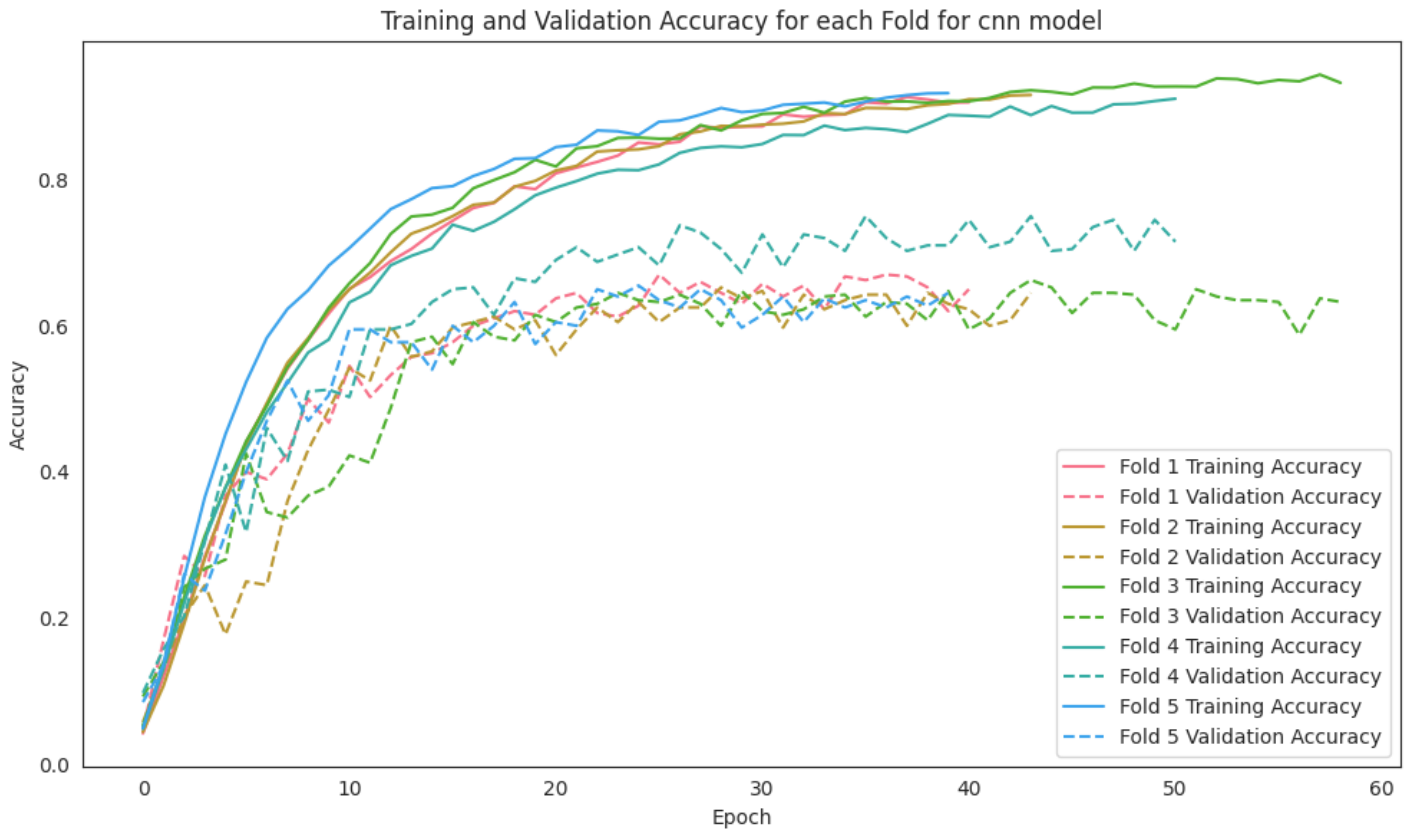
- Training and Validation sets using a **cross-validation** approach
- 5 folds
- **Original dataset**: (80% / 20% split)
  Train samples: 1600
  Val samples: 400
- **Augmented dataset**:(94% / 6% split)

  Train samples: 6400
  Val samples: 400
- **No augmentation** is applied to the **validation** set to ensure that the evaluation process remains unbiased



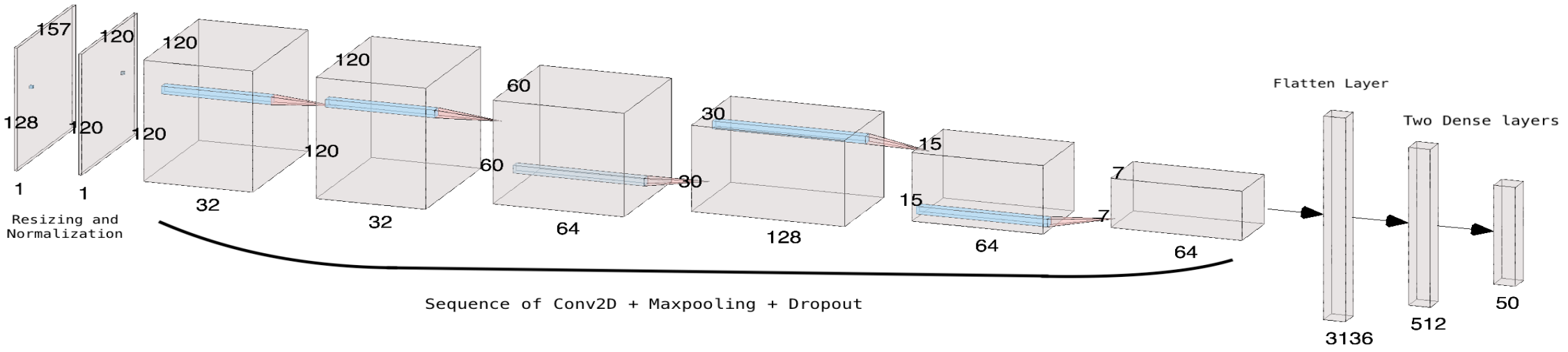
Training Strategy

- Optimizer: Adam, learning_rate=0.001
- Sparse Categorical Crossentropy loss
- Standard Scaler Normalization
- 100 epochs
- early stopping
- model checkpoint callback



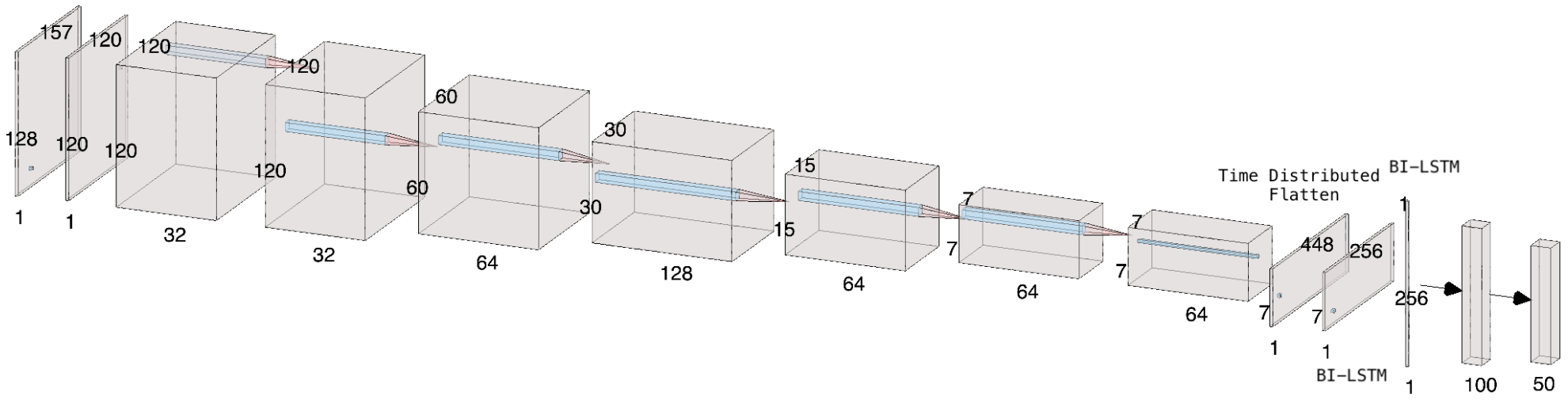
CNN Model

- 3x3 kernels , ReLU activations, "same" padding
- **Dropout** layers with a rate of 0.2 and 0.5
- **Batch Normalization** is employed after deeper convolutional layers



CNN-LSTM

- CNN layers as previous
- Output of CNN is **permutated** to set the time dimension for the LSTM
- Couple of **Bidirectional LSTM** with 128 units
- Dropout of 0.25



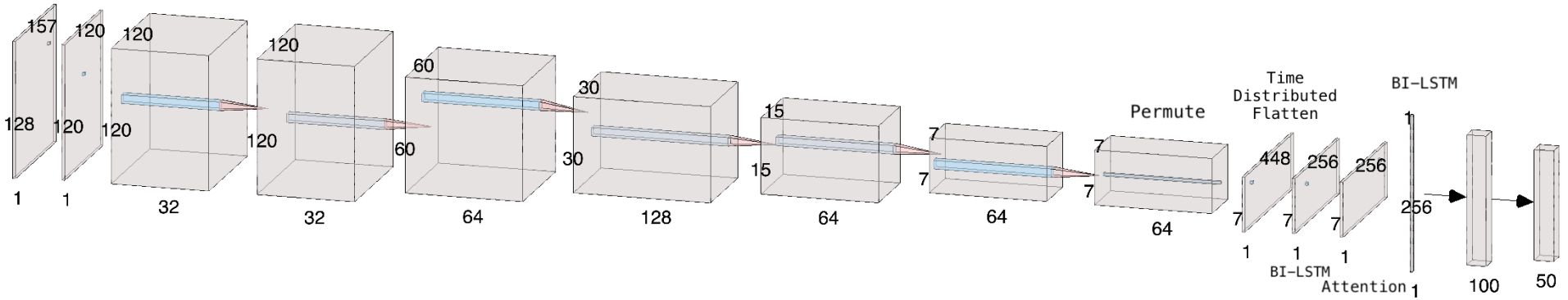
CNN-LSTM-Attention

- CNN-LSTM as previous
- Simple attention mechanism between the two LSTM layers
- To weigh the importance of different time steps within the sequence

$$Attention(\mathbf{X}, \mathbf{W}, \mathbf{b}) = softmax(\tanh(\mathbf{XW} + \mathbf{b}))$$

C. Raffel and D. P. W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," 2016.

X [timestep, LSTM units]
W [LSTM units, 1]
B [timestep, 1]

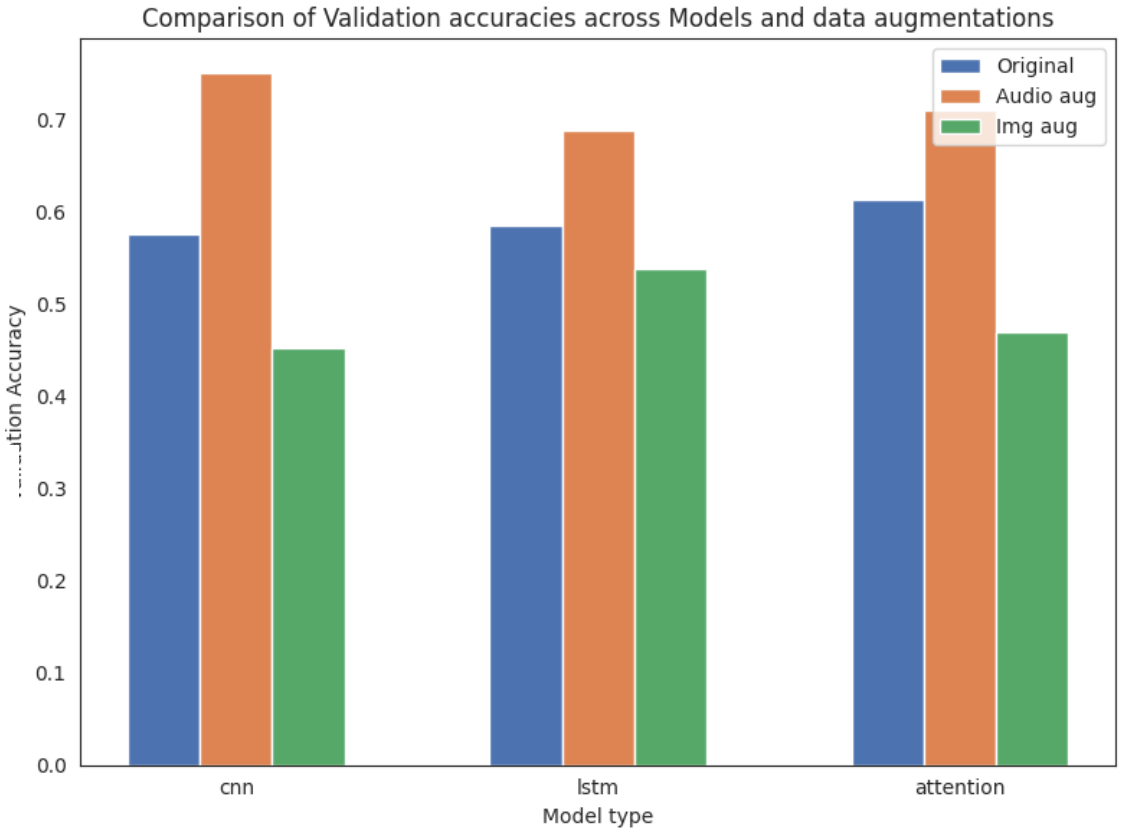


Model Evaluation

- CNN has the **best accuracy**
- **Audio** augmentations are **crucial**
- **Image** augmentations **worse** the accuracy in **every** model

	original	audio	img	params
cnn	0.57	**0.75**	0.45	2 M
lstm	0.58	**0.69**	0.54	1.4 M
attention	0.61	**0.71**	0.47	1.4 M

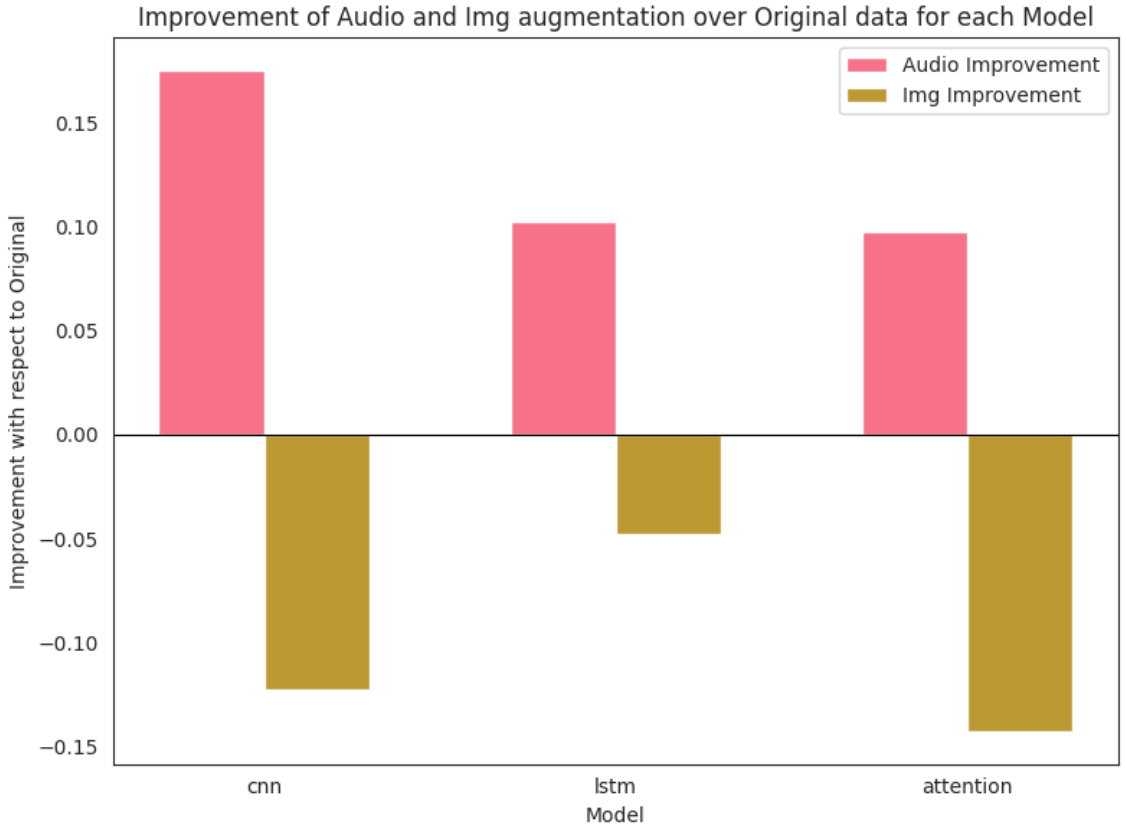
Human Accuracy is 81%



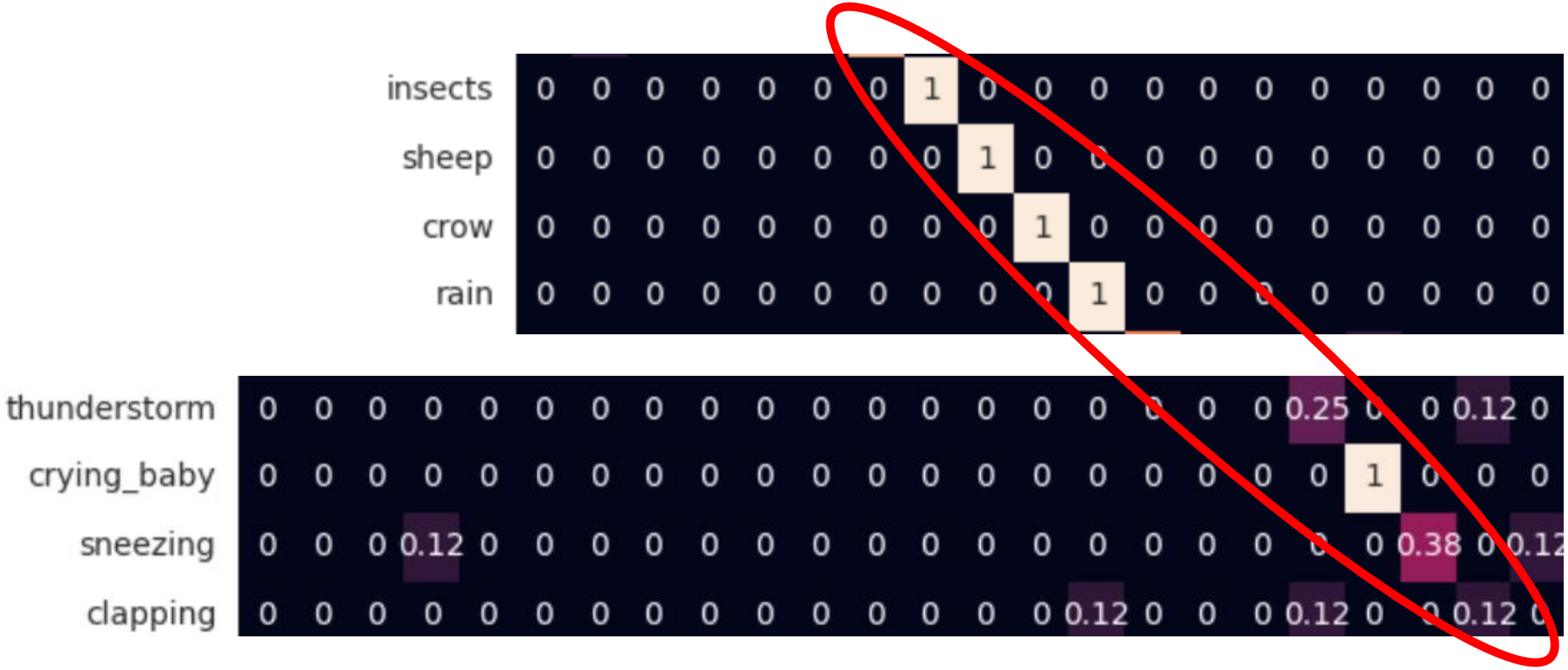
Augmentations analysis

- Augmentations **types** could have strong influence
- Adjusting Dataset **size** could help

	original	audio	img	params
cnn	0.57	**0.75**	0.45	2 M
lstm	0.58	**0.69**	0.54	1.4 M
attention	0.61	**0.71**	0.47	1.4 M



Confusion Matrix



Conclusions

- Implemented and compared three different **deep learning architectures**
- Three different **data preparation** techniques
- Importance of **audio augmentations**
- Minimal increase of only 200 parameters led to a significant 2% **improvement** in accuracy

Future works:

- Implementing more sophisticated **attention models,** such as multi-head attention or transformer based architectures like SOTA ones
- Exploring **better** augmentation **strategies**

	original	audio	img	params
cnn	0.57	**0.75**	0.45	2 M
lstm	0.58	**0.69**	0.54	1.4 M
attention	0.61	**0.71**	0.47	1.4 M