**Learning Strategies for Voice Disorder Detection**

Preprocessing-

1. Silence at the start and end of the audio recordings is removed (if exists).
2. Down sampling to 16 kHz.
3. Each file is segmented into multiple 500 ms long snippets, with a 400 ms overlap of subsequent snippets.
4. Dataset augmentation – each sample is pitch-shifted by 8 half-semitones up and 8 half-semitones down.

Input Representation-

1. **Feature based representation with MFCC – for SVM and AE models**

Each snippet is divided into multiple blocks for a Short-time Fourier transform (STFT) with a block size of 512 and a hop size of 128 samples, respectively. Then, for each block, 20 MFCCs are extracted to form a 20 \* 63 input matrix and their mean values and standard deviations over blocks are calculated, resulting in a 40-dimensional feature vector per snippet.

1. **Mel-spectrogram – for CNN models**