Spoken Language Identification(SLID)

The process of detecting language from an audio clip by an unknown speaker, regardless of gender, manner of speaking, and distinct age speaker, is defined as spoken language identification (SLID). The considerable task is to recognize the features that can distinguish between languages clearly and efficiently. The model uses audio files and converts those files into spectrogram images. It applies the convolutional neural network (CNN) to bring out main attributes or features to detect output easily.

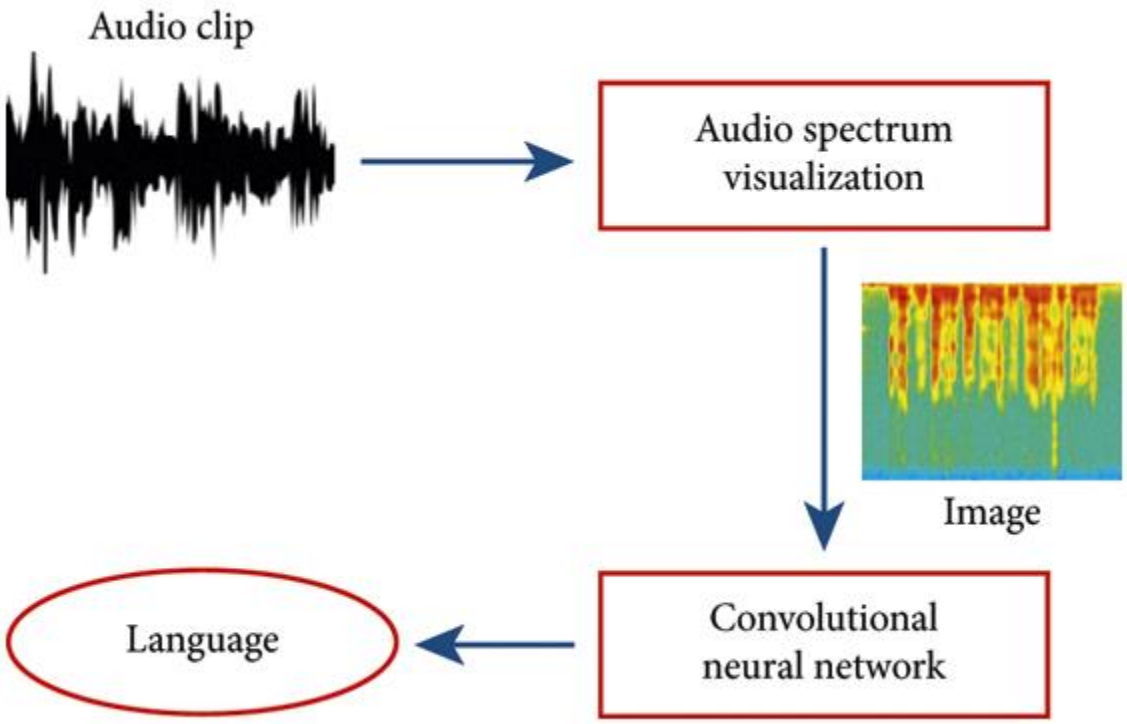
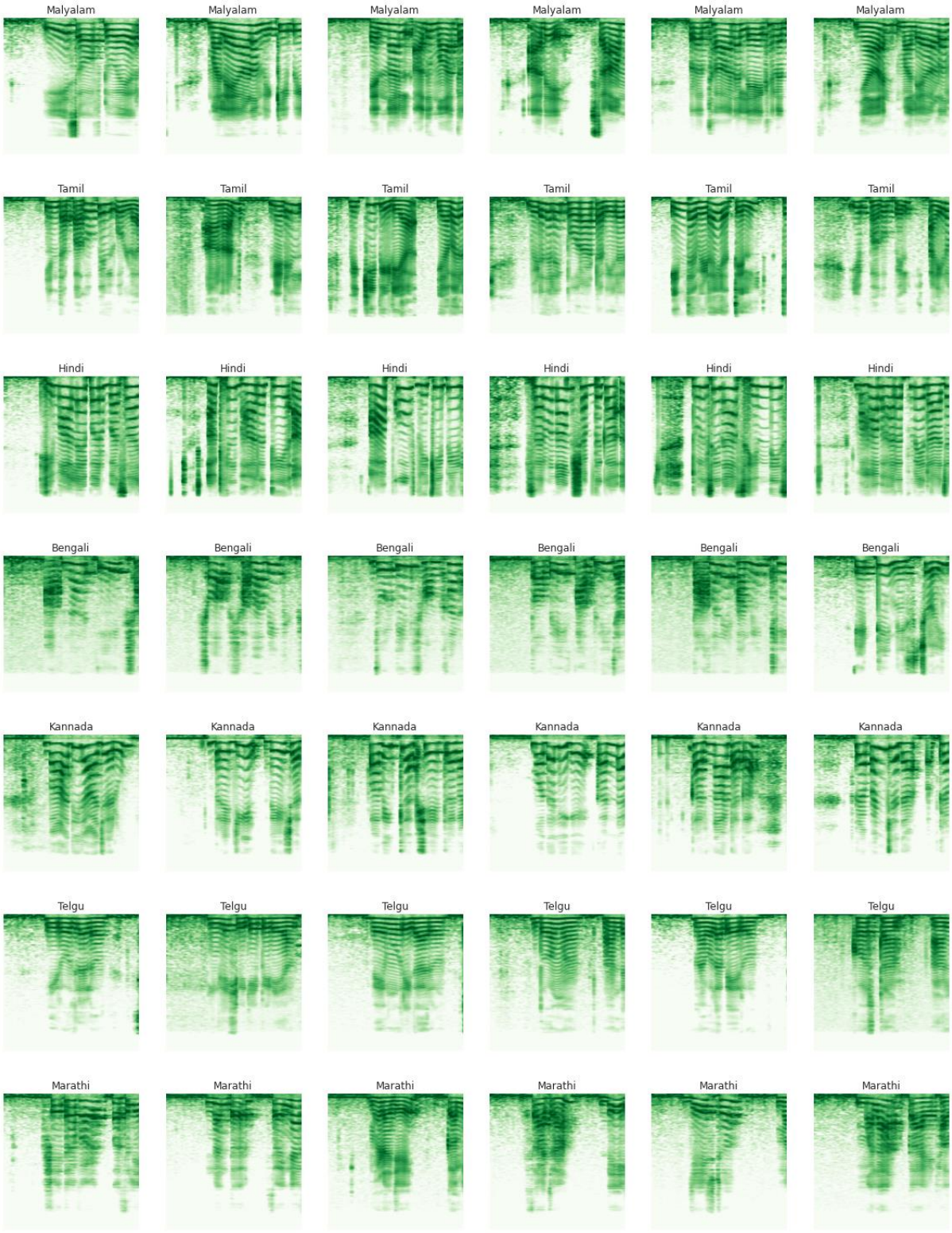


Fig: Workflow of a SLID Model.

Mel-Spectrograms of Different Indic Languages.



Convolutional Neural Network Architecture

```
Model: "LID_Model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_70 (Conv2D)           (None, 196, 196, 64)      128
_________________________________________________________________
max_pooling2d_46 (MaxPooling (None, 98, 98, 64)        0
_________________________________________________________________
conv2d_71 (Conv2D)           (None, 98, 98, 128)       8320
_________________________________________________________________
max_pooling2d_47 (MaxPooling (None, 49, 49, 128)       0
_________________________________________________________________
conv2d_72 (Conv2D)           (None, 49, 49, 128)       16512
_________________________________________________________________
flatten_23 (Flatten)         (None, 307328)            0
_________________________________________________________________
dense_46 (Dense)             (None, 64)                19669056
_________________________________________________________________
dense_47 (Dense)             (None, 7)                 455
=================================================================
Total params: 19,694,471
Trainable params: 19,694,471
Non-trainable params: 0
_________________________________________________________________
```

Idea: The main idea behind a model is to build Mel-Spectrogram Images for uploading audio files and use a trained CNN model to determine the type of Dialect Spoken by the Speaker in that audio files.

References: https://www.hindawi.com/journals/cin/2021/5123671/