CNN structure:

Input (Depends on type whether Spectrogram or Topographical Map)
|
|
Conv2D -> ReLU -> MaxPooling -> BatchNorm -> Dropout
|
|
Conv2D -> ReLU -> MaxPooling -> BatchNorm -> Dropout
|
|
Flatten
|
|
Dense -> ReLU -> Dropout
|
|
Output Layer (e.g., Dense with sigmoid activation for binary classification)

We can add LSTM layers as well.(before flatten)

There are many other standard Networks which can be modified to use for our purpose .
1. AlexNet
2. GoogleNet

These are just the general overview we may work deep to see what can be improved to achieve maximum accuracy.