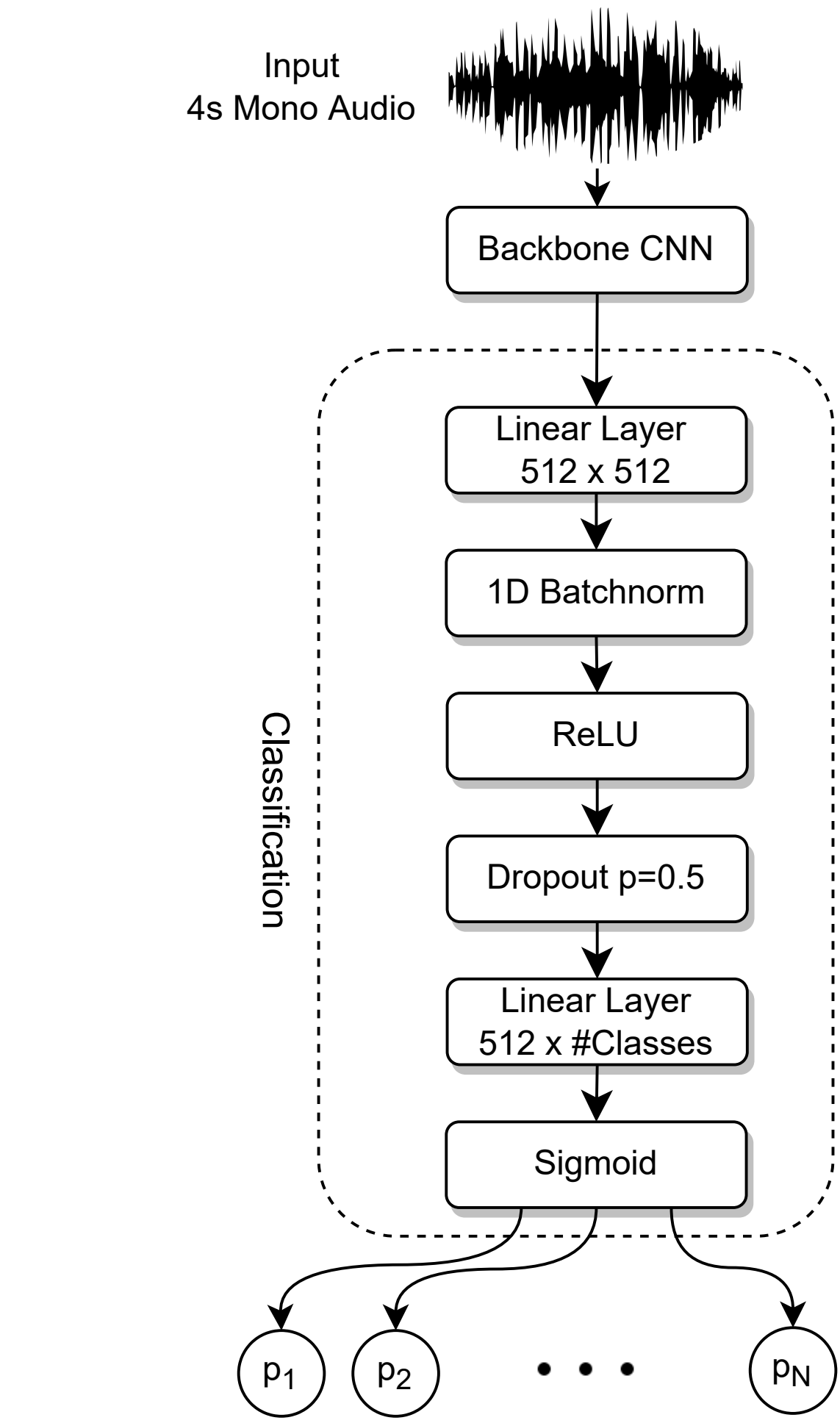
Input
4s Mono Audio
Backbone CNN
Classification
Linear Layer
512 x 512
1D Batchnorm
ReLU
Dropout p=0.5
Linear Layer
512 x #Classes
Sigmoid
p1
p2
pN