Database RAVDESS

- The RAVDESS consists of total 1440 samples recorded from 24 professional actors (12 male and 12 female).
- It encompasses eight emotions: calm, surprise, neutral, happy, angry, sad, fearful, and disgust
- The sampling rate of original speech signals is 16000Hz.
- We have considered the 5 sec long speech samples to keep the uniformity in the Mel frequency log spectrogram.
- If the sample length is less than 5 sec then it is padded to 5 sec long. Otherwise the samples are cropped to 5 sec.

Sr. No.	Emotion	Number of Samples
1	Angry	192
2	Calm	192
3	Disgust	192
4	Fear	192
5	Happy	192
6	Neutral	96
7	Sad	192
8	Surprise	192

Data	Samples
Total samples	1440
Train (70%)	1005
Test (15 %)	218
Validation (15%)	217

Flow Diagram –MFLS+DCNN

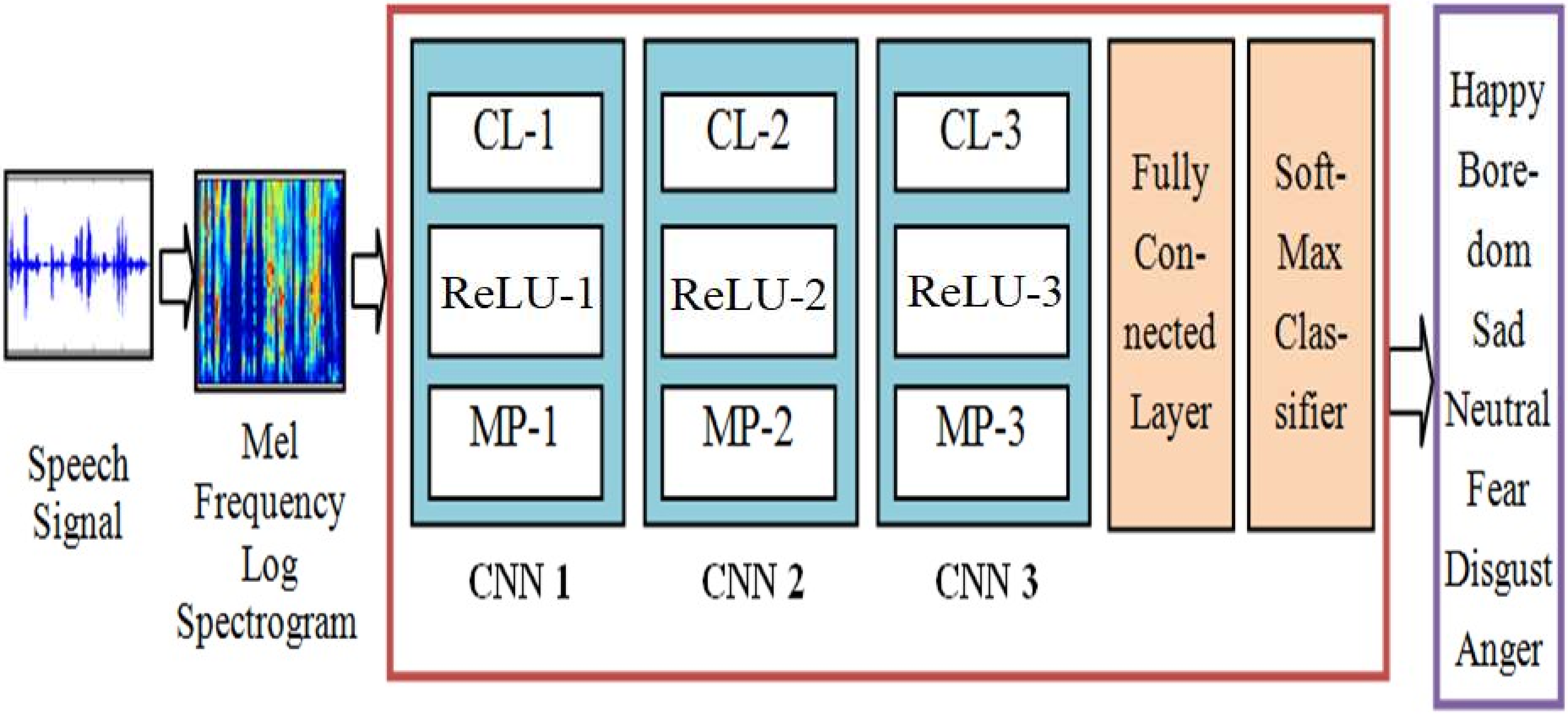
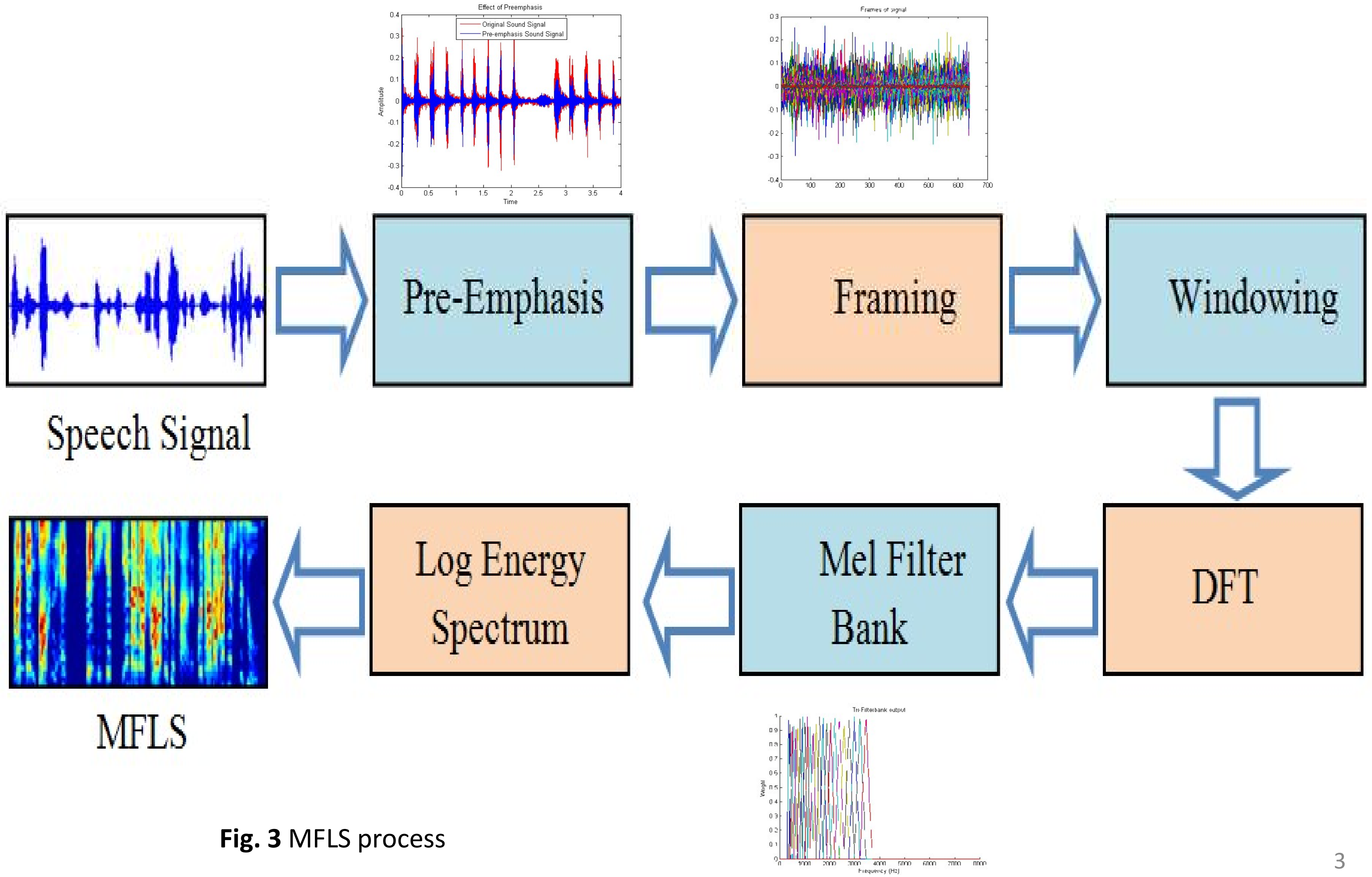


Fig. 2 Process flow of SER based on MFLS+DCNN

Mel Frequency Logarithmic Spectrum (MFLS)



Speech Signal → Pre-Emphasis → Framing → Windowing → DFT → Mel Filter Bank → Log Energy Spectrum → MFLS

Fig. 3 MFLS process

Mel Frequency Logarithmic Spectrum (MFLS)

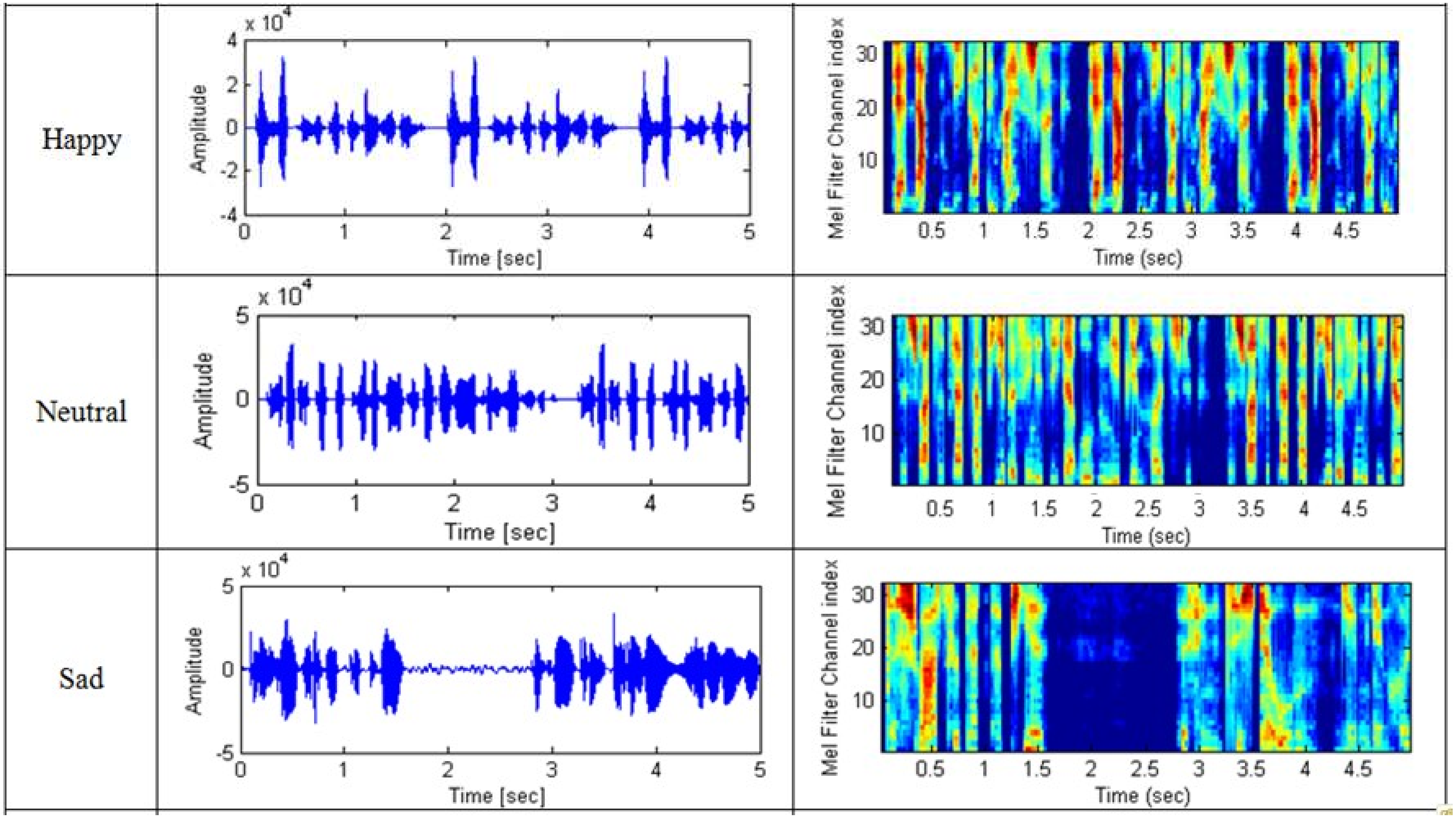
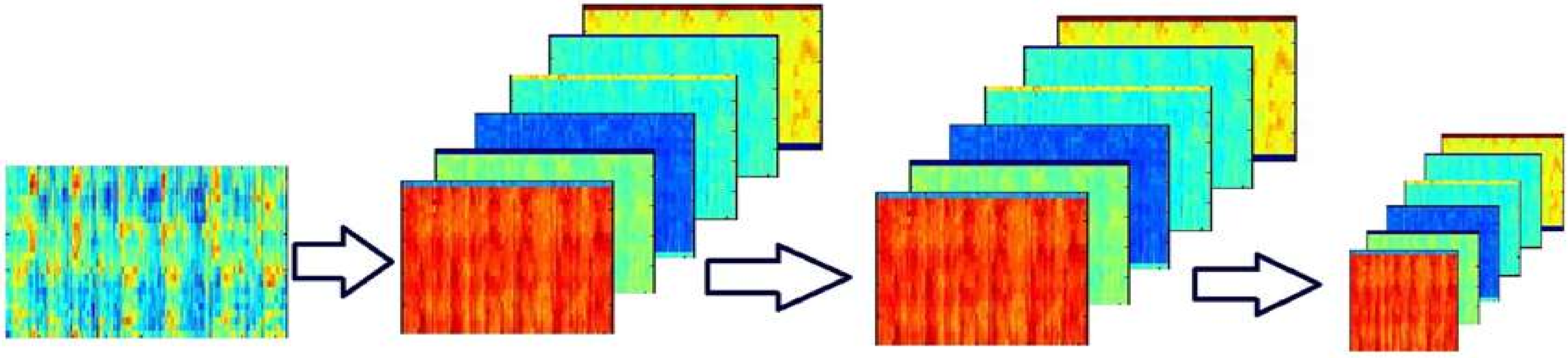


Fig. 4 MFLS representation

Convolution Neural Network Architecture



(32x249)
Mel Filter
Energy Signal

(32x249x6)
Convolution Layer 1

(32x249x6)
ReLULayer 1

(16x124x6)
Max Pooling Layer 1

Fig. 5 CNN architecture

❑ Training Method

Mini-Batch Gradient Descent Algorithm is used for training CNN.

In mini-batch gradient descent algorithm n complete dataset is divided in to small batches b, then the model coefficients are updated using model error.

$$E_t[f(w)] = \frac{1}{b}\sum_{i=(t-1)b+1}^{tb} f(w, x_i) \qquad (1)$$

The weights are updated using equation;

$$w^{t+1} = w^t - \mu\nabla_w E_t[f(w^t)] \qquad (2)$$

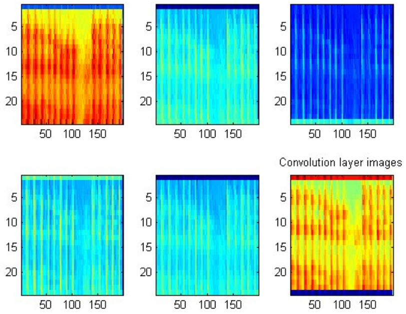
Where,
Et = model error , Xi = training samples
W =weights of filter kernel
F =cost function
μ= Learning rate
∇= Gradient of cost function
tb = number of batches

CNN Architecture Layer

❑ Convolution layer

$$I_{CNN} = \mathrm{Im} \otimes W_{N\times N} \quad (3)$$

Convolution layer gives local connectivity of image local region and discriminancy weights of image region.

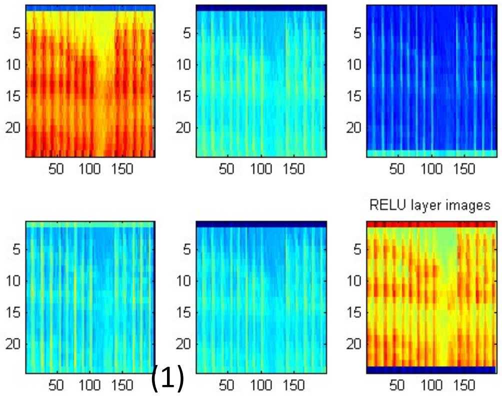


❑ ReLU Layer

$$I_{\mathrm{ReLU}} = \begin{cases} 0 & if(I_{CNN(i,j)} < 0) \\ I_{CNN(i,j)} & Otherwise \end{cases}$$

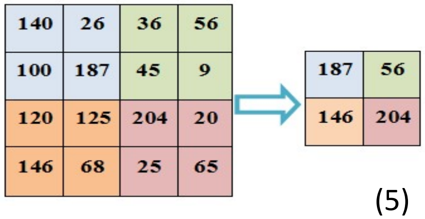
(4)

Rectified linear Unit removes the linearity by replacing negative weights by zero.



(1)

❑ Max Pooling layer

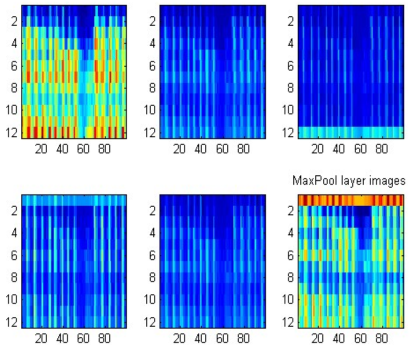


140	26	36	56
100	187	45	9
120	125	204	20
146	68	25	65

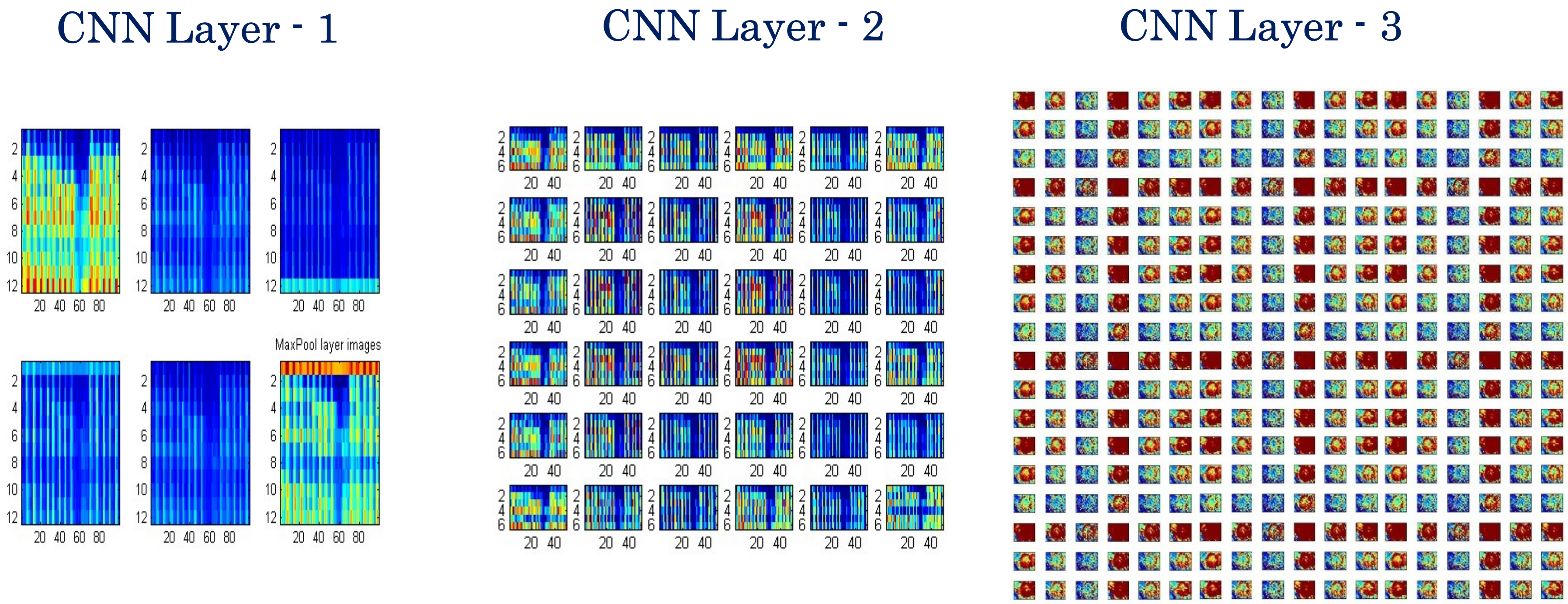
187	56
146	204

(5)

Max Pooling layer down sample the image to reduce feature map.



CNN layer Feature Visualization (2-D Color)



CNN Layer - 1

CNN Layer - 2

CNN Layer - 3

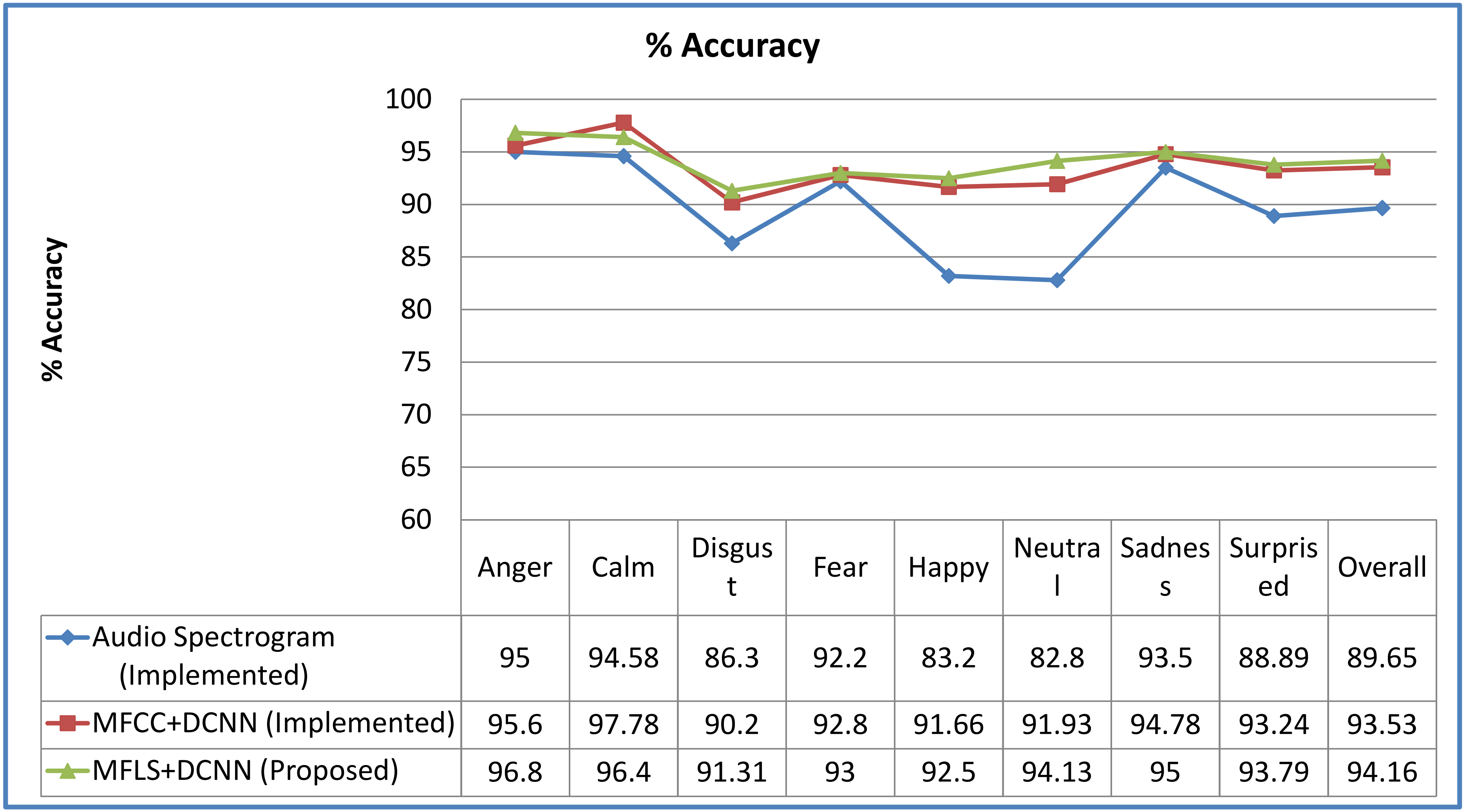
Fig. 6 Visualization of various DCNN maps

Experimental Results

Table 3 : Feature maps for various CNN layers

Layer	Sub-Layer	Kernel Size	Stride	Feature map
Input Layer	Mel Frequency Log Spectrogram	-	-	32 × 249
CNN layer 1	Convolution Layer 1	3 × 3 × 6	1	29 × 247 × 6
	ReLU Layer 1	-	-	29 × 247 × 6
	Max Pooling Layer 1	2 × 2	2	14 × 123 × 6
CNN layer 2	Convolution Layer 2	3 × 3 × 36	1	12 × 121 × 36
	ReLU Layer 2	-	-	12 × 121 × 36
	Max Pooling Layer 2	2 × 2	2	6 × 60 × 36
CNN layer 3	Convolution Layer 3	3 × 3 × 216	1	4 × 58 × 216
	ReLU Layer 3	-	-	4 × 58 × 216
	Max Pooling Layer 3	2 × 2	2	2 × 29 × 216
FC Layer	-	-	-	**12528 × 1**

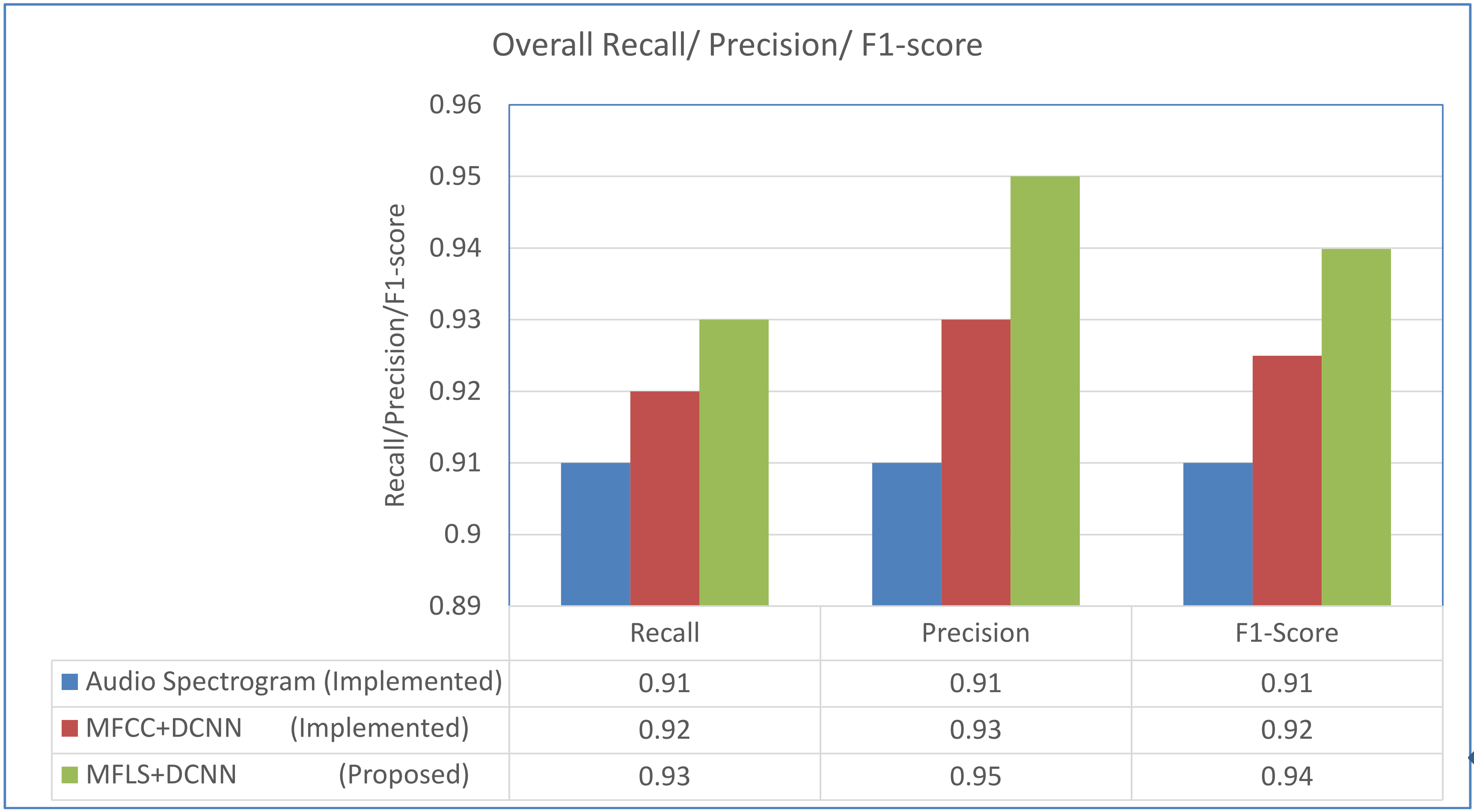
Accuracy for RAVDESS dataset



	Anger	Calm	Disgust	Fear	Happy	Neutral	Sadness	Surprised	Overall
◆ Audio Spectrogram (Implemented)	95	94.58	86.3	92.2	83.2	82.8	93.5	88.89	89.65
■ MFCC+DCNN (Implemented)	95.6	97.78	90.2	92.8	91.66	91.93	94.78	93.24	93.53
▲ MFLS+DCNN (Proposed)	96.8	96.4	91.31	93	92.5	94.13	95	93.79	94.16

Fig. 9 % Accuracy for MFLS+MFCC database

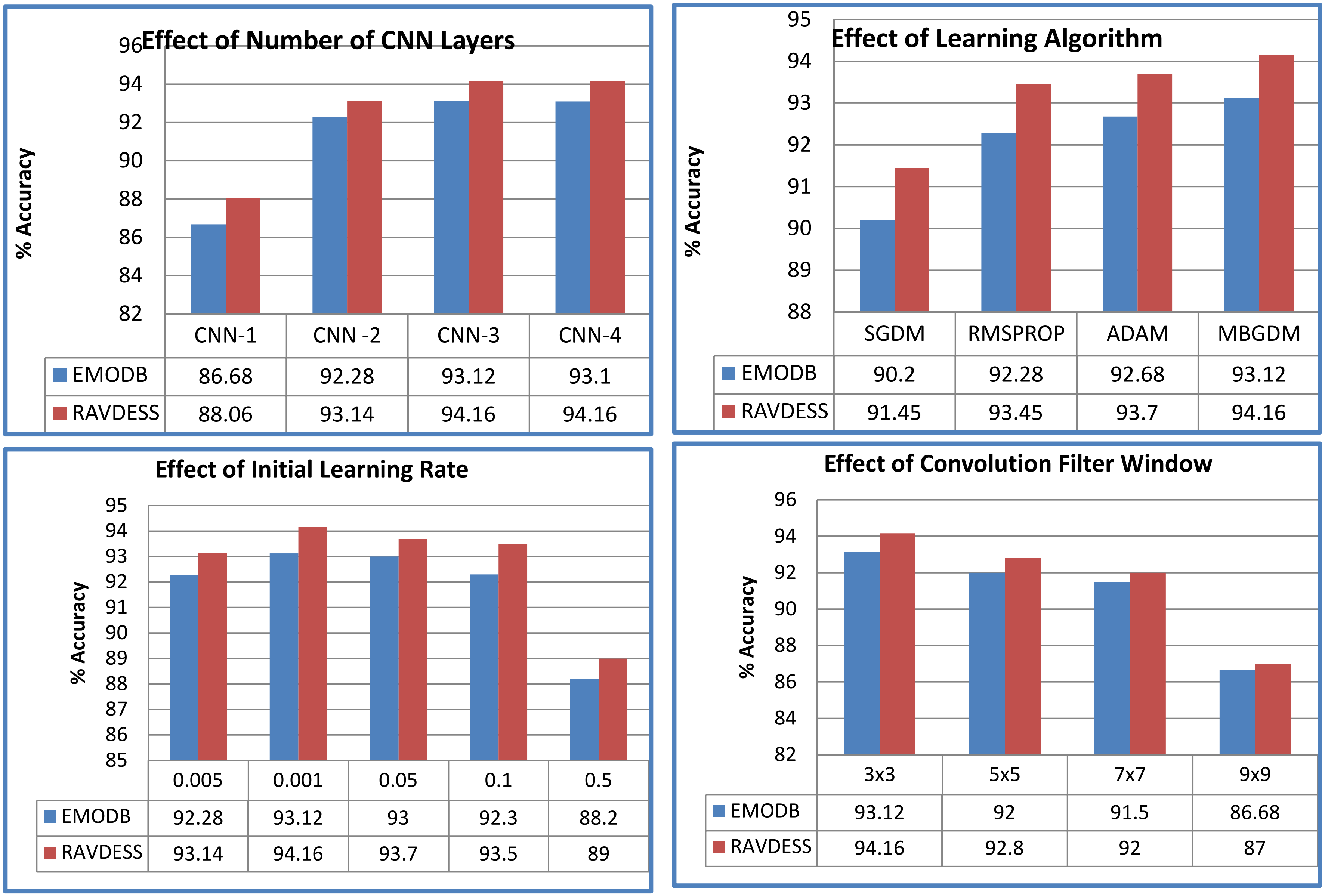
Recall, Precision and F1-score for RAVDESS dataset



	Recall	Precision	F1-Score
🟦 Audio Spectrogram (Implemented)	0.91	0.91	0.91
🟥 MFCC+DCNN (Implemented)	0.92	0.93	0.92
🟩 MFLS+DCNN (Proposed)	0.93	0.95	0.94

Fig 10. Recall/ Precision/F1-score for MFLS+DCNN for RAVDESS database

Performance for Various Parameters of DCNN



Effect of Number of CNN Layers

	CNN-1	CNN -2	CNN-3	CNN-4
🟦 EMODB	86.68	92.28	93.12	93.1
🟥 RAVDESS	88.06	93.14	94.16	94.16

Effect of Learning Algorithm

	SGDM	RMSPROP	ADAM	MBGDM
🟦 EMODB	90.2	92.28	92.68	93.12
🟥 RAVDESS	91.45	93.45	93.7	94.16

Effect of Initial Learning Rate

	0.005	0.001	0.05	0.1	0.5
🟦 EMODB	92.28	93.12	93	92.3	88.2
🟥 RAVDESS	93.14	94.16	93.7	93.5	89

Effect of Convolution Filter Window

	3x3	5x5	7x7	9x9
🟦 EMODB	93.12	92	91.5	86.68
🟥 RAVDESS	94.16	92.8	92	87

Fig. 11 % Accuracy for various DCNN parameters