UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Audio & Image classification, Image retrieval

Progetto finale
Digital Signal & Image Management

Giacomo Stoffa - 830159 Leonardo Riva - 830647 Riccardo Merlo - 829805

Audio classification

Dataset

Task: identificare chi sta parlando, tra i componenti del gruppo
L'interlocutore deve pronunciare 3 cifre (senza altri vincoli)

Costruzione dataset:

registrazione cifre (più volte)

0, 1, 2, 3, 4, 5, 6, 7, 8, 9

→

raggruppate casualmente per il train

419, 655, 803, …

→

test set registrato manualmente

train set di 100 osservazioni (+ 20 di validation); test set di 33 osservazioni

Modelli

1. NN senza features pre-calcolate
2. NN con features audio base
3. NN con features da modello pre-trainato

↓ complessità features crescente

Per tutti i modelli, sono state scelte 3 sigmoidi, rispetto ad 1 softmax

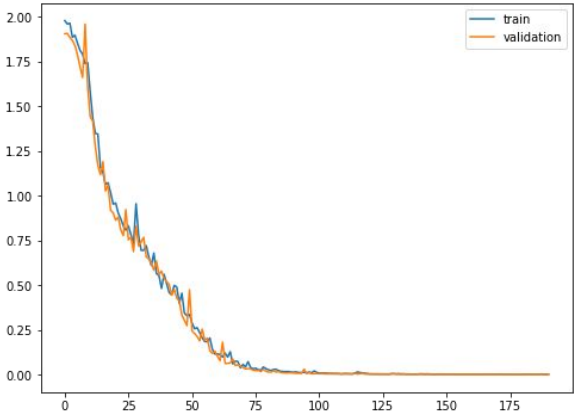
→ si riesce a discriminare meglio quando a parlare non è nessuno tra i componenti

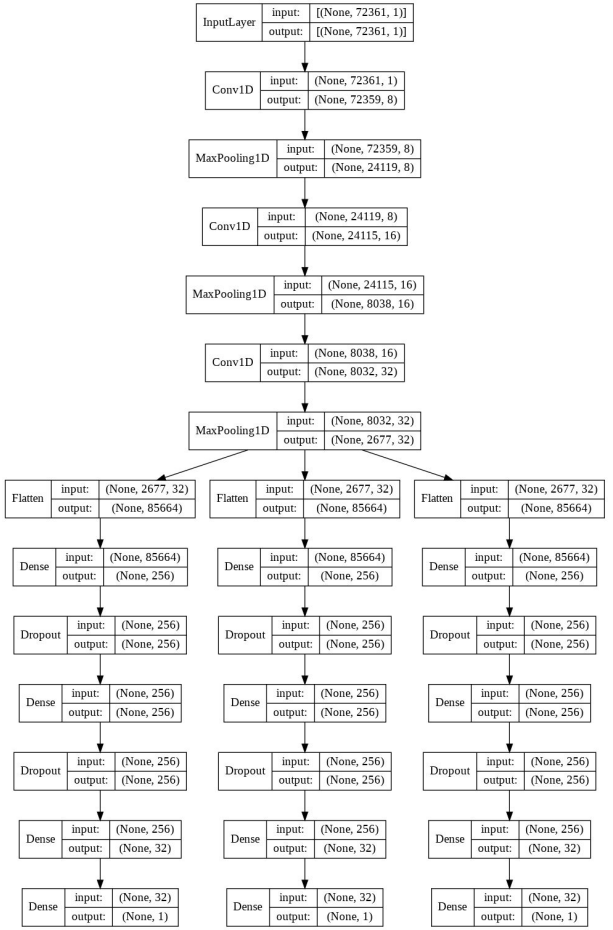
→ si ottengono risultati migliori

Modello 1 (NN senza features pre-calcolate)

2 fasi:

- Convoluzioni e MaxPooling (a 1 dimensione)
- Flatten, Dense e Dropout

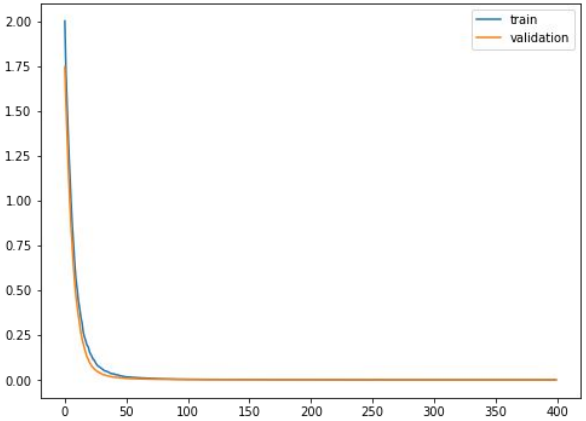


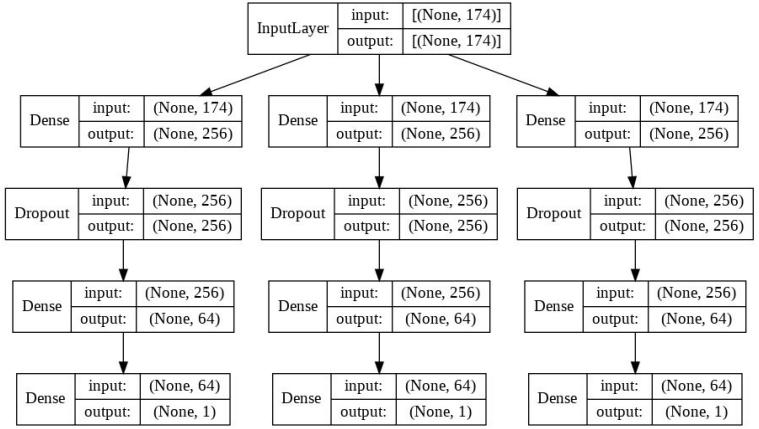


Modello 2 (NN con features audio base)

Librosa features → utilizzate quelle che hanno dimensione indipendente da lunghezza audio

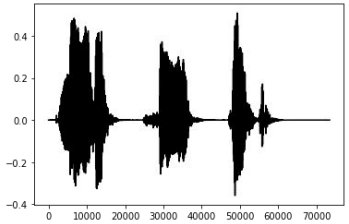
- [mfcc, chroma_stft, melspectrogram, spectral_contrast, tonnetz, energy]



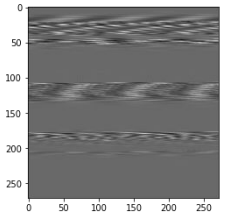


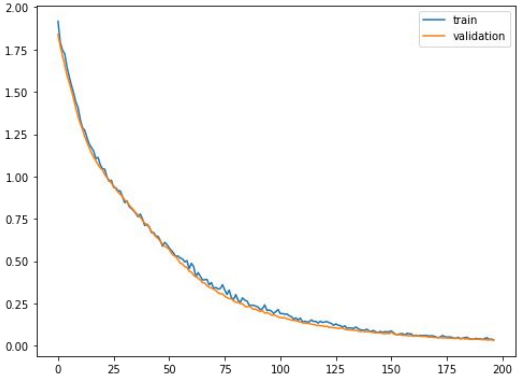
Modello 3 (NN con features da modello pre-trainato)

segnale 1D

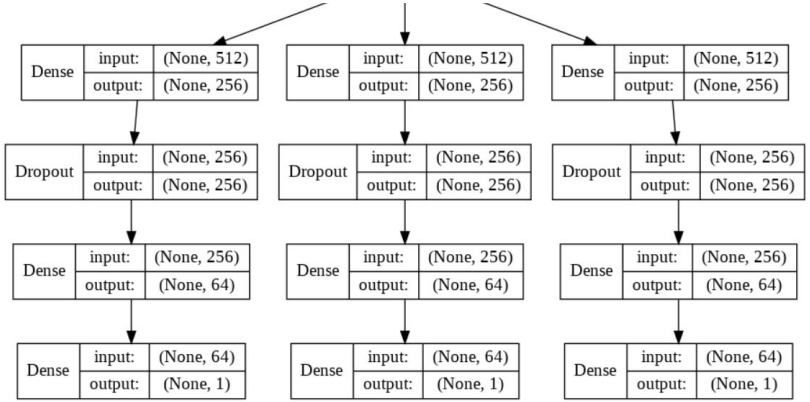


reshape 2D e resize





VGG16 freezed model

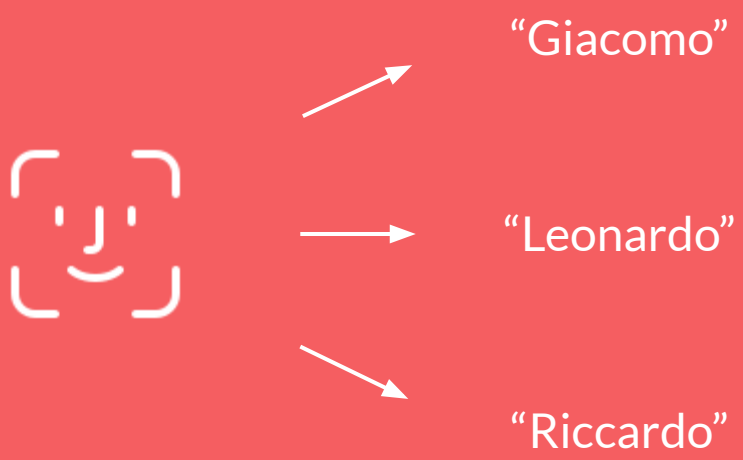


Risultati

F1-score	Modello 1	Modello 2	Modello 3
Giacomo	0.17	0.91	0.85
Leonardo	0.51	0.95	0.59
Riccardo	0.53	0.87	0.78
	0.40	**0.91**	**0.74**

Tempo di training	480s	130s	420s

Image classification



Workflow

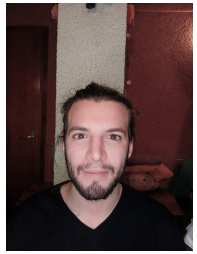
Task: identificare il volto di una persona , tra i componenti del gruppo.
Assegnare probabilità predizione classificatore.

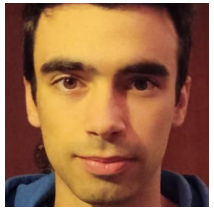
Costruzione Dataset:

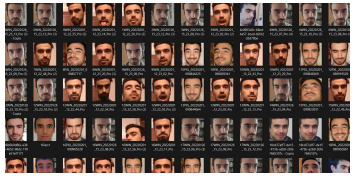
Scattate foto per ogni componente da diverse posizioni e angolazioni → Utilizzo di un Face Detector per ritagliare i volti (pulizia supervisionata) → Partizionamento in train, validation e test

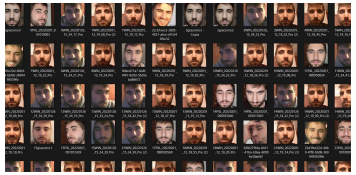
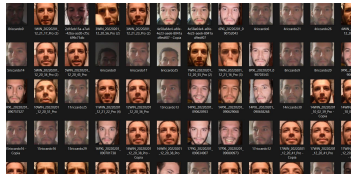
Modello reti neurali MobileNetv2 → Modello reti neurali VGG16 → Modello LBPH (Local Binary Pattern Histogram)

Dataset







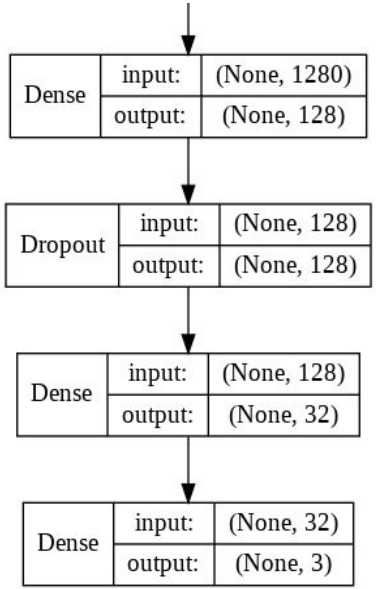



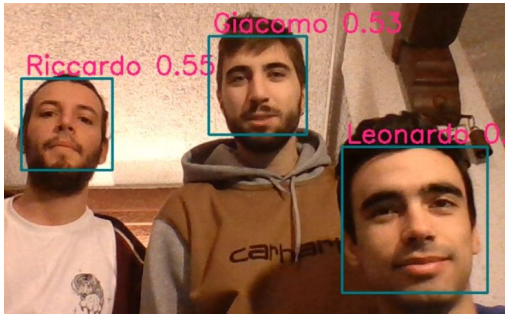
Train: 672 images

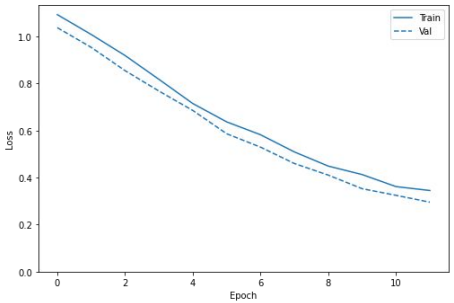
Validation: 73 images

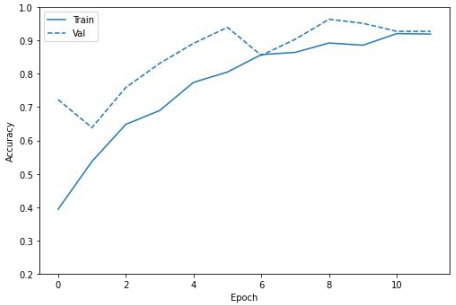
Test: 83 images

Modello 1 (MobileNetV2)

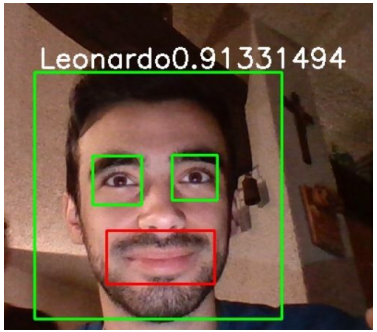




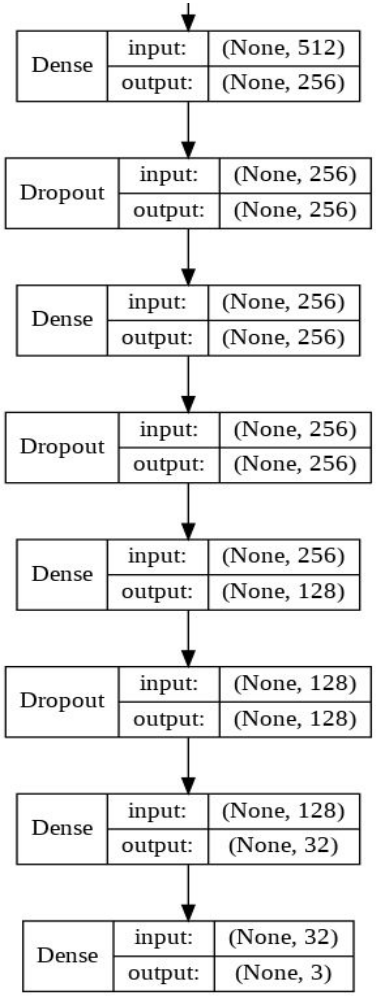




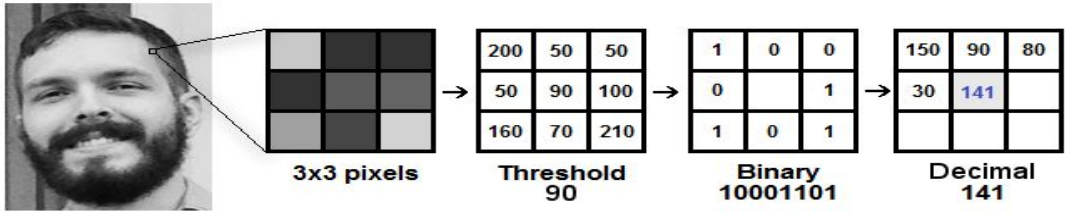
Modello 2 (vgg16)

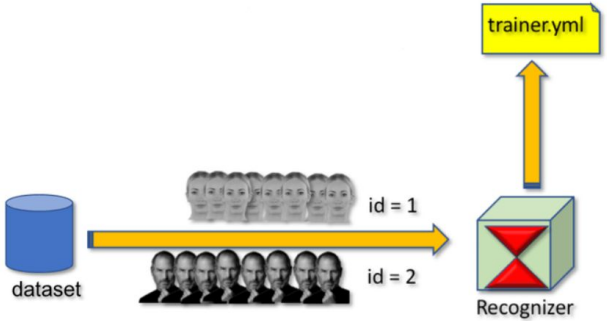


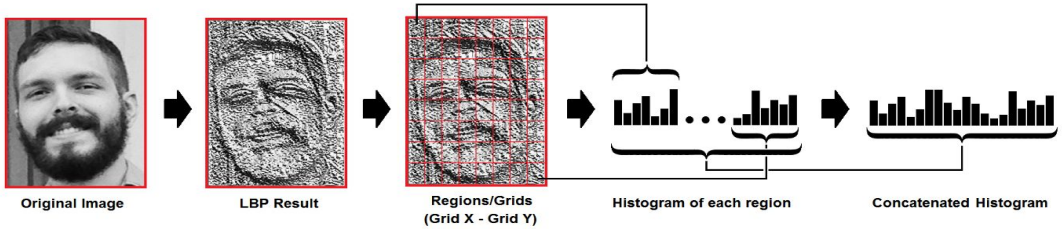
	precision	recall	f1-score	support
0	0.92	1.00	0.96	23
1	1.00	0.94	0.97	32
2	0.96	0.96	0.96	28
accuracy			0.96	83
macro avg	0.96	0.97	0.96	83
weighted avg	0.97	0.96	0.96	83

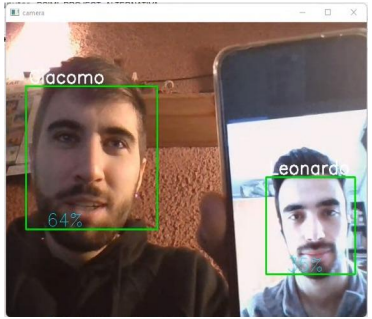


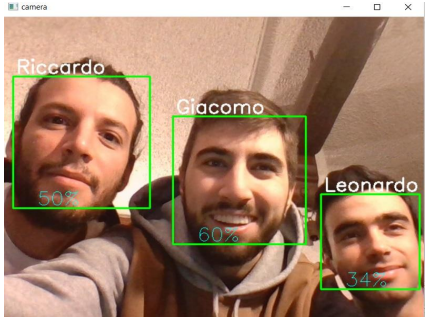
Modello 3 (LBPH)











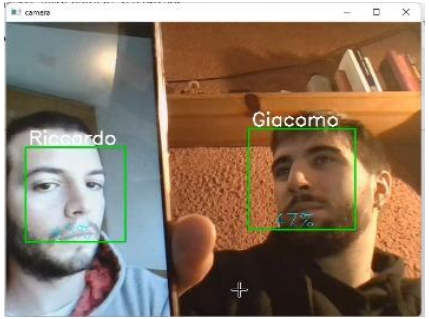


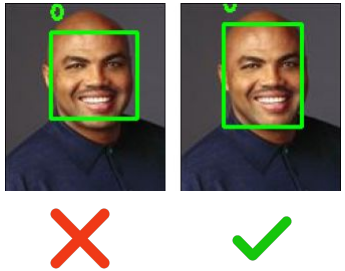
Image retrieval

Dataset

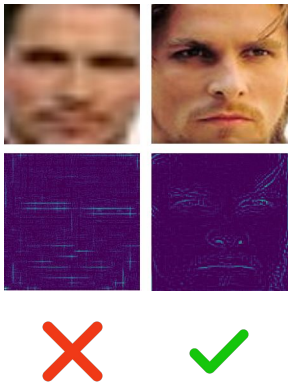
Task: trovare le 10 facce più simili, tra quelle presenti nel dataset, all'immagine in input

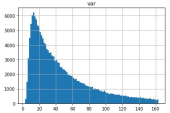
Costruzione dataset:

face detection tramite MTCNN



rimozione immagini low quality





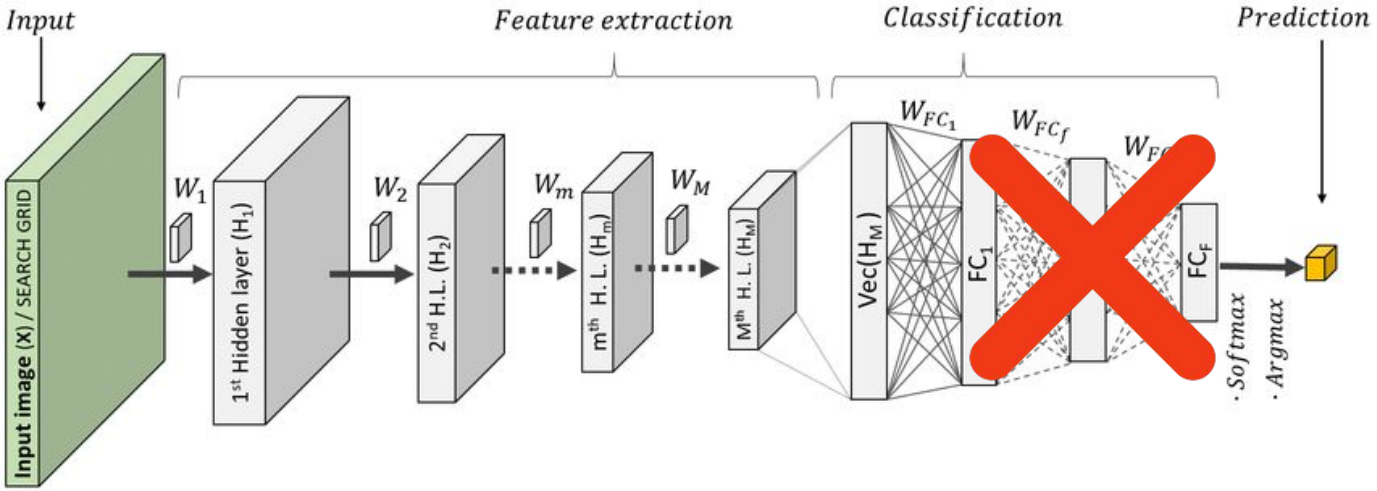
estrazione feature

$[[\ x_1, x_2, x_3, x_4, x_5, \ldots\ ]]$

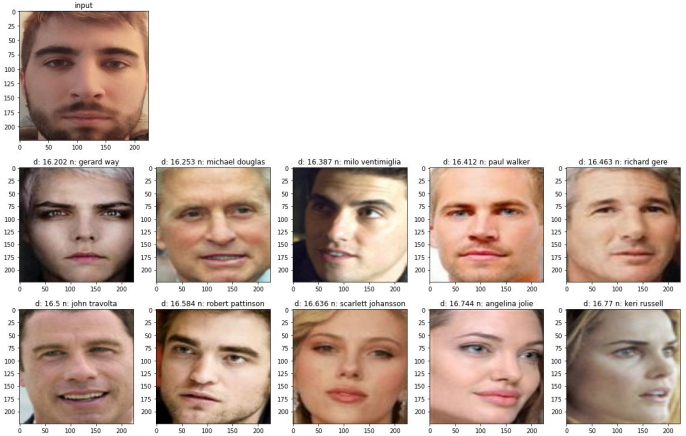
Modelli

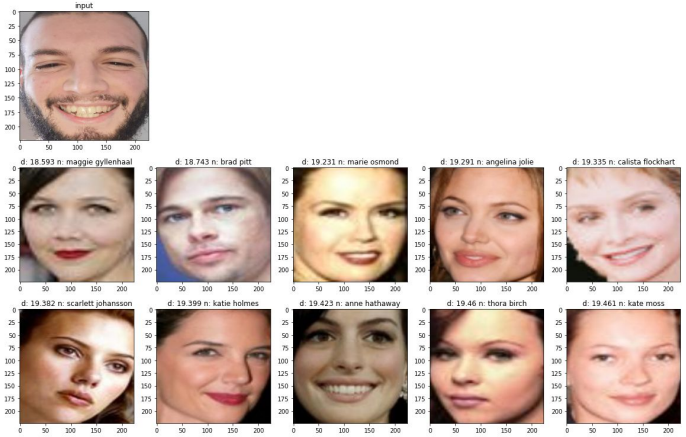
- Mobilenet V2
- DenseNet
- VGGFACE2 con ResNet50

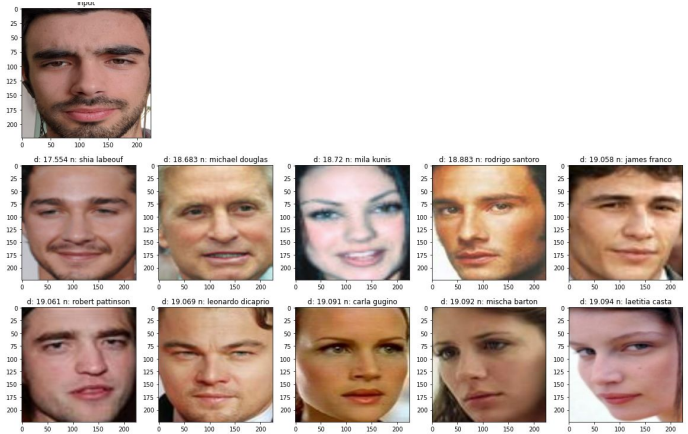
- *pesi Imagenet*
- *pesi VggFace*



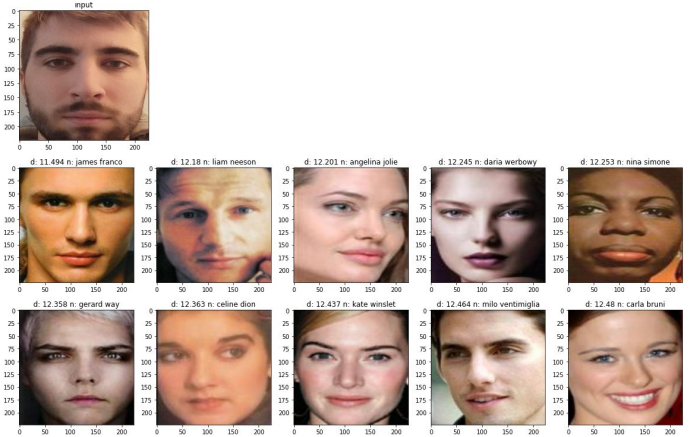
Modello 1 (MobileNetV2)

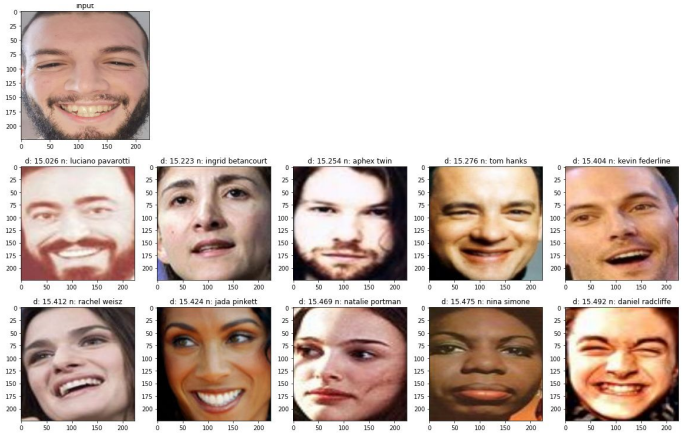


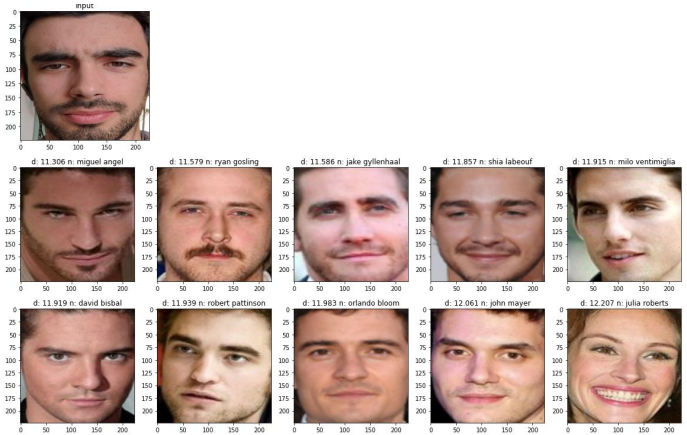




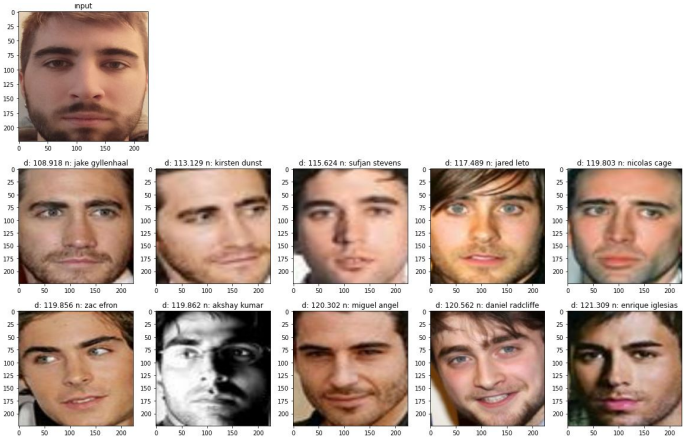
Modello 2 (DenseNet)

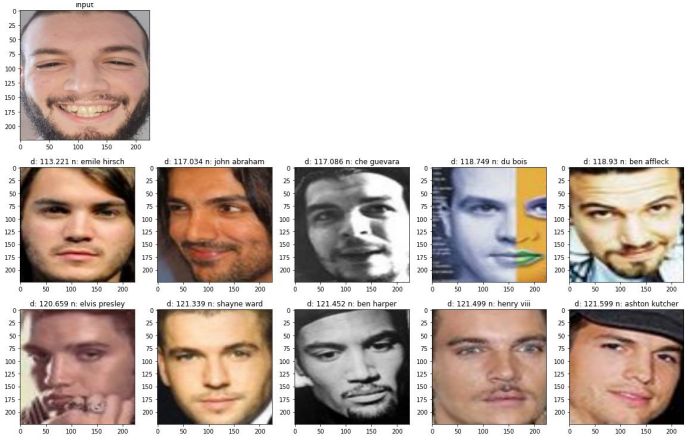


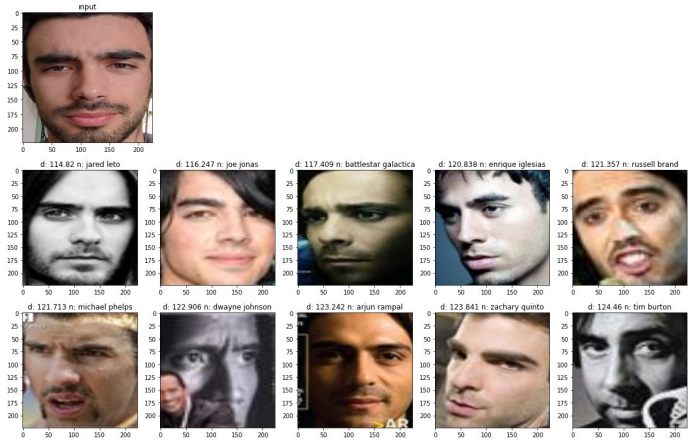




Modello 3 (VggFace2)







THE END