**Scientific background**

Lie detection via video leverages involuntary facial expressions, eye movements, and body gestures potentially indicative of deception. 3D Convolutional Neural Networks (3D CNNs) excel in this area by processing spatial and temporal data together, capturing transient expressions and movements critical in identifying lies. They analyze motion and expressions across frames, detecting patterns and inconsistencies. On the other hand, CNN-LSTMs combine CNNs' spatial feature extraction with LSTMs' ability to manage long sequences, making them adept at recognizing patterns and contextual behaviors over time.

**Preprocessing**

The videos are fed into a preprocessing pipeline that is outlined in the following major steps:

1. Extracting the subject’s face from the Images in the video.
2. Sampling the extracted faces sequence into a fixed size through down sampling (frame skipping) to get sequences of similar length to feed to the neural networks. (sampling is done after face cropping because not all frames successfully yield a face image after cropping)
3. Resizing the final sequences to a fixed width and height to ensure the input shape is consistent across all samples.

**Model Training**

After preprocessing, the sampled and normalized images are fed into a 3D CNN network to extract features from the videos (3D CNNs are particularly good here compared to CNN-LSTMs for lie detection as they can pick up on features related to motion in videos more which can allow the model to understand different gestures and facial movements that might be correlated with lying), finally these features are fed to fully connected layers to further process the data and discover more hidden relations, and at the end there is a fully connected layer followed by one neuron in the output layer that outputs the probability of deception in the input video.

In this approach, feature extraction is done in the 3D CNN layers.

**Important remarks and discoveries**

On using a random data split (this causes unreal results because certain subjects appear so many times in the dataset so the models can potentially learn to identify these subjects rather than detect lies) The accuracy yielded is very high reaching ~92% specially using a CNN-LSTM model, although it’s a lower with a 3D CNN at around 83%, but surprisingly when manually splitting the data to ensure that the test set contains videos of subjects that have never appeared in the train set (to eliminate all bias superstitions) the CNN-LSTM achieves poor results with accuracy below 60% while the 3D CNN maintains decent results with peek accuracy of around 83% on both train and test sets (same as the automatic, biased split) So we chose to continue with the 3D CNN as the preferred model when classifying the videos for visual cues that indicate lying.