**Video Classification Using CNN+LSTM**

**Objective:**

Classify videos (5 classes in this study) using CNN and LSTM with two models – ConvLSTM & Long Term Recurrent Convolutional Network (LRCN).

***Reference:***[*https://bleedai.com/human-activity-recognition-using-tensorflow-cnn-lstm/*](https://bleedai.com/human-activity-recognition-using-tensorflow-cnn-lstm/)

**Process:**

- **Step 1:** Download and Visualize the Data with its Labels

- **Step 2:** Pre-process the Dataset

- **Step 2.1:** Frame Extraction

- **Step 2.2:** One Hot Encoding of Labels

- **Step 3:** Split the Data into Train and Test Set

- **Step 4:** Implement the ConvLSTM Approach

- **Step 4.1:** Construct the Model

- **Step 4.2:** Compile & Train the Model

- **Step 4.3:** Plot Model’s Loss & Accuracy Curves

- **Step 5:** Implement the LRCN Approach

- **Step 5.1:** Construct the Model

- **Step 5.2:** Compile & Train the Model

- **Step 5.3:** Plot Model’s Loss & Accuracy Curves

- **Step 6:** Test the Best Performing Model on Test Videos

Video Extraction 🡪 Frame Extraction (input of No. of Frames and Sequences per video – same for every video) 🡪 Feeding to Model 🡪 Label Classification

1. Video Extraction – UFC50 Dataset

* **50** Action Categories
* **25** Groups of Videos per Action Category
* **133** Average Videos per Action Category
* **199** Average Number of Frames per Video
* **320** Average Frames Width per Video
* **240** Average Frames Height per Video
* **26** Average Frames Per Seconds per Video

We are taking only 5 out of 50 categories as training labels for this example.

Baseball Pitch

Biking,

Diving,

High Jump,

Horse Race

Resize every video – 64 x 64

1. Frame Extraction (input of No. of Frames and Sequences per video – same for every video)

Provide Sequence Length (No. of Frames to Extract from a video)

(How it works:

E.g. –

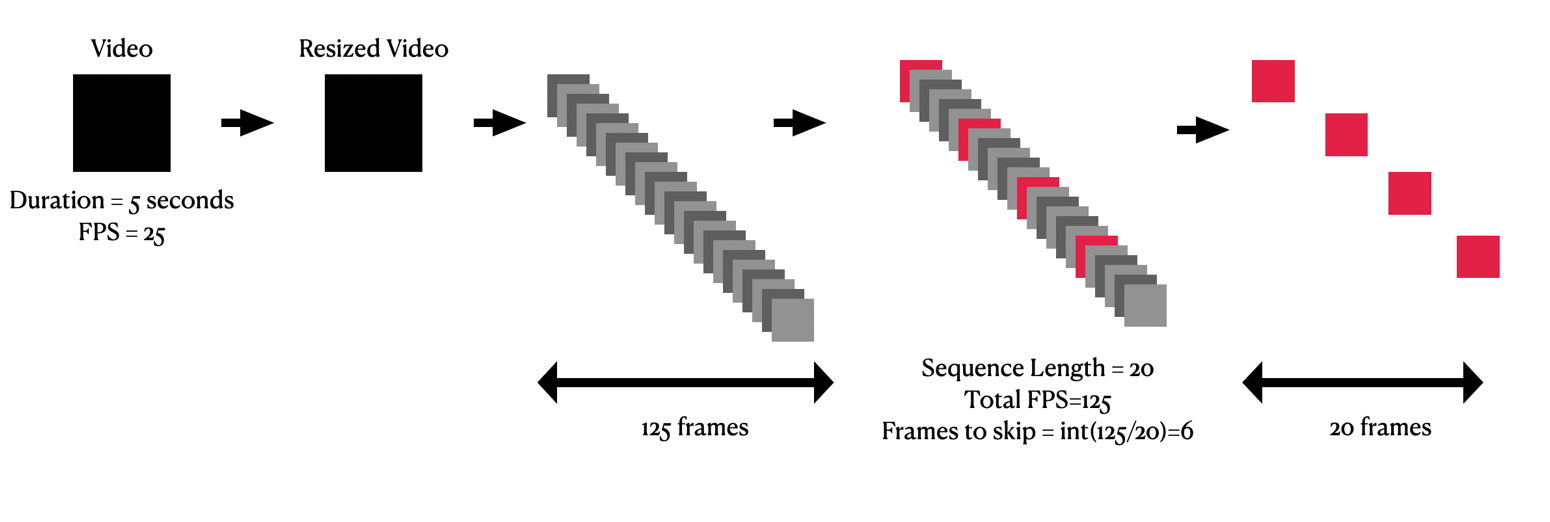
5 second video with 25 fps.

Total Frames = 25\*5 = 125

Sequence Length = 20

No. of Frames to skip = INT(125/20) = 6

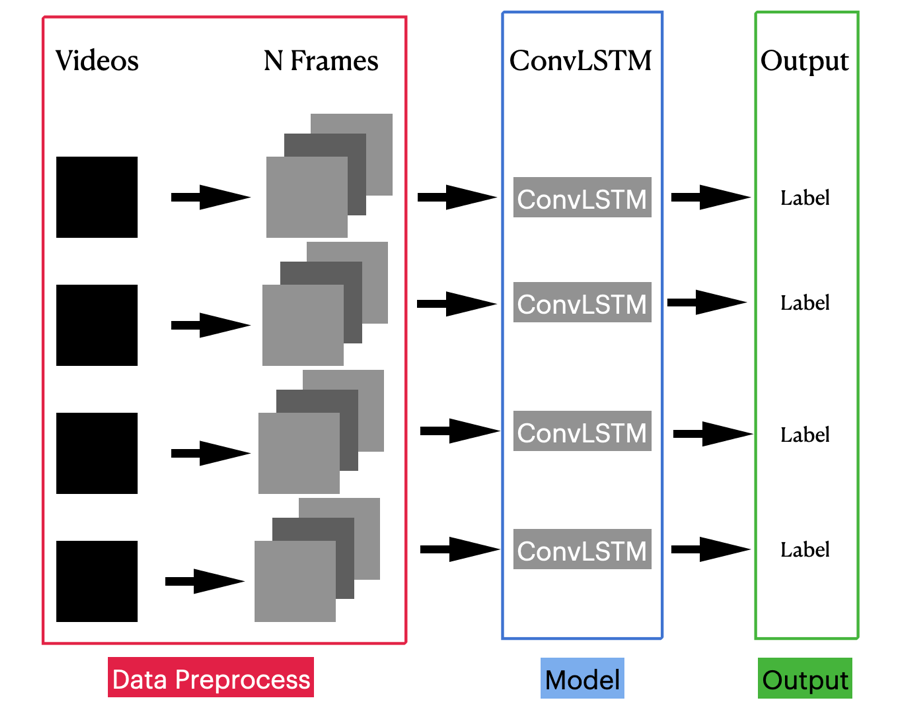
Every 6th frame is extracted in a sequence of the video, gathering a total of 20 frames)

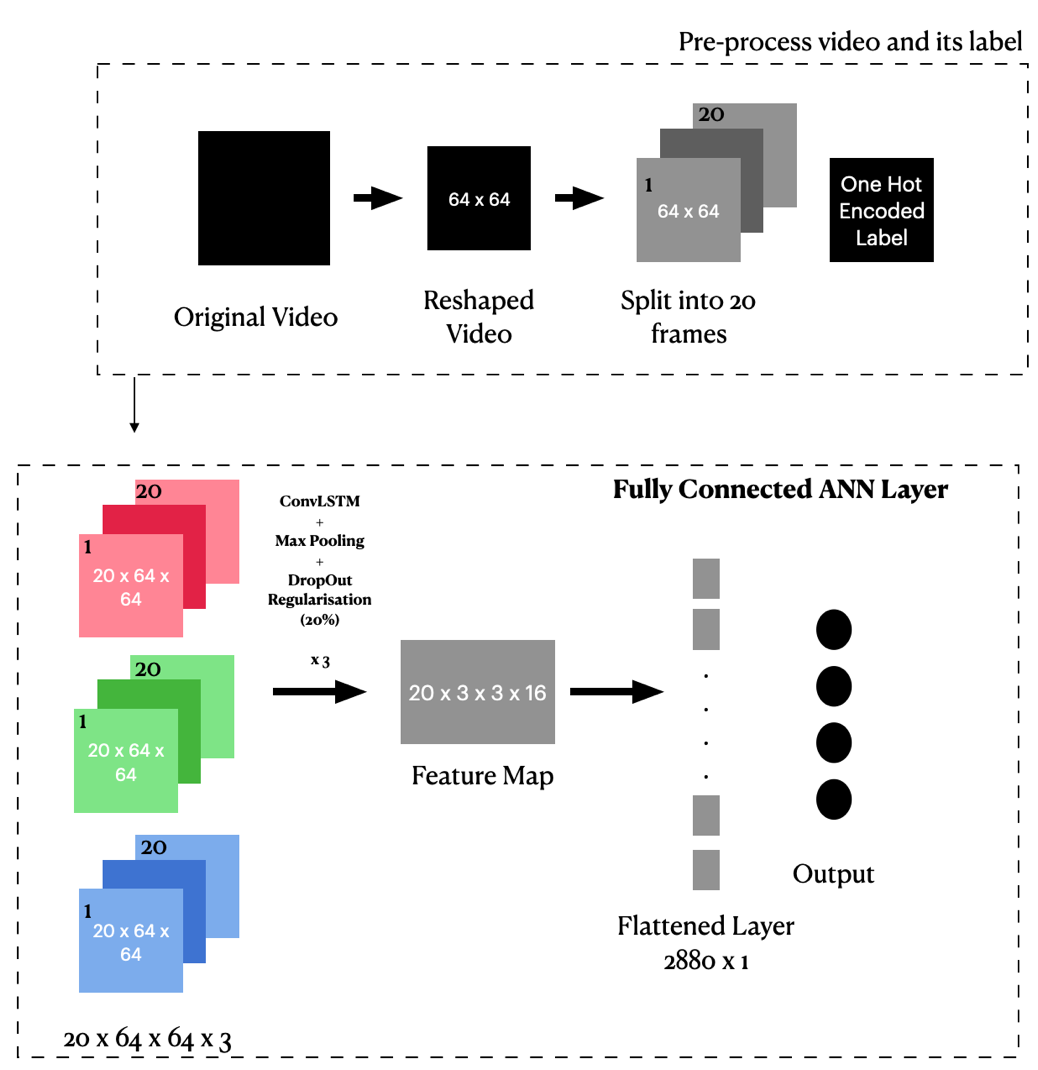


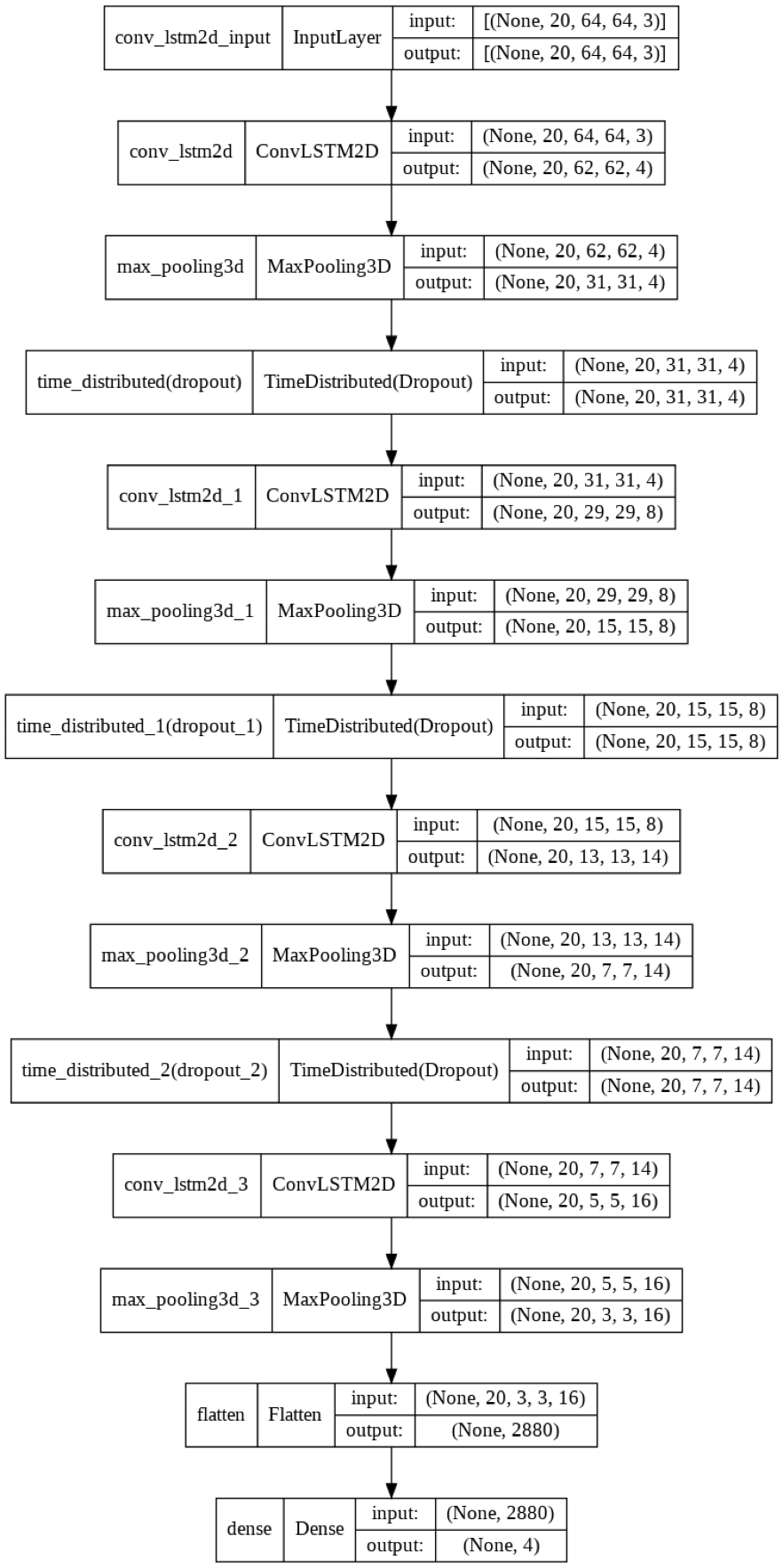
1. Feeding to Model

**Model 1: ConvLSTM**

**Architecture:**

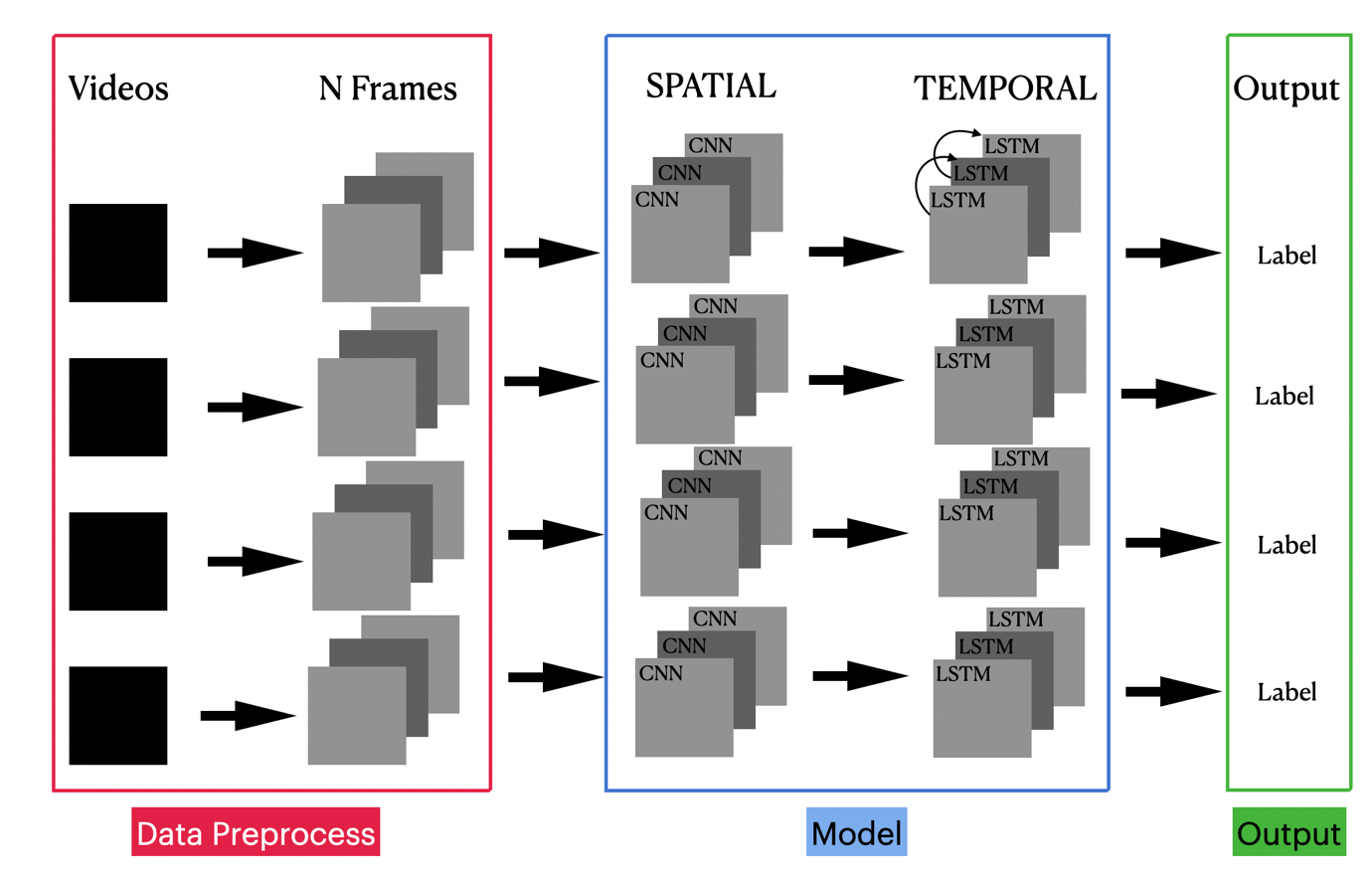






**Model 2: LRCN**

**Architecture:**

****

