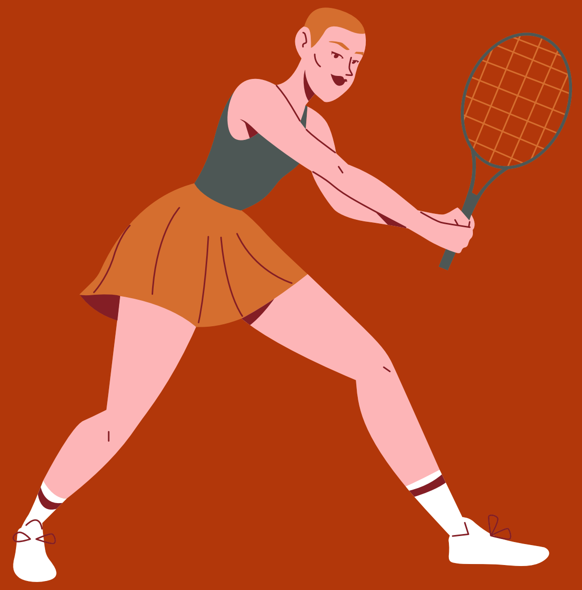
Tennis Shots Recognition



through Human-Pose estimation and Deep LSTM-based Neural Network

Cecilia Assolito 1857897

Antonello Giorgio 1836529

VISION AND PERCEPTION

SAPIENZA
UNIVERSITÀ DI ROMA

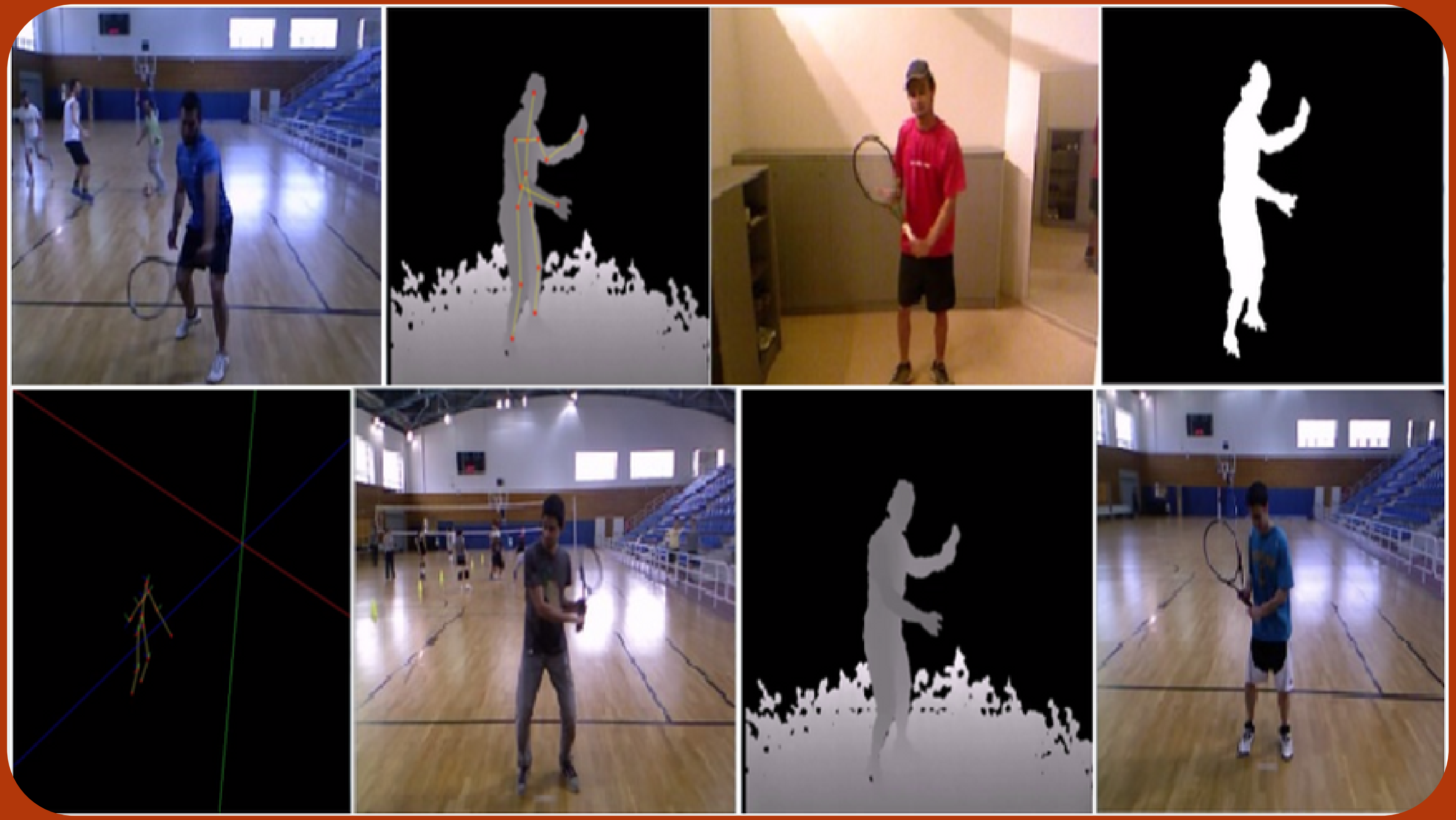
Goal of the project

Implement a Deep Neural Network model to classify different types of typical tennis strokes, starting with input in video format.

Dataset

THETIS (THree dimEnsional TennIs Shots [1])

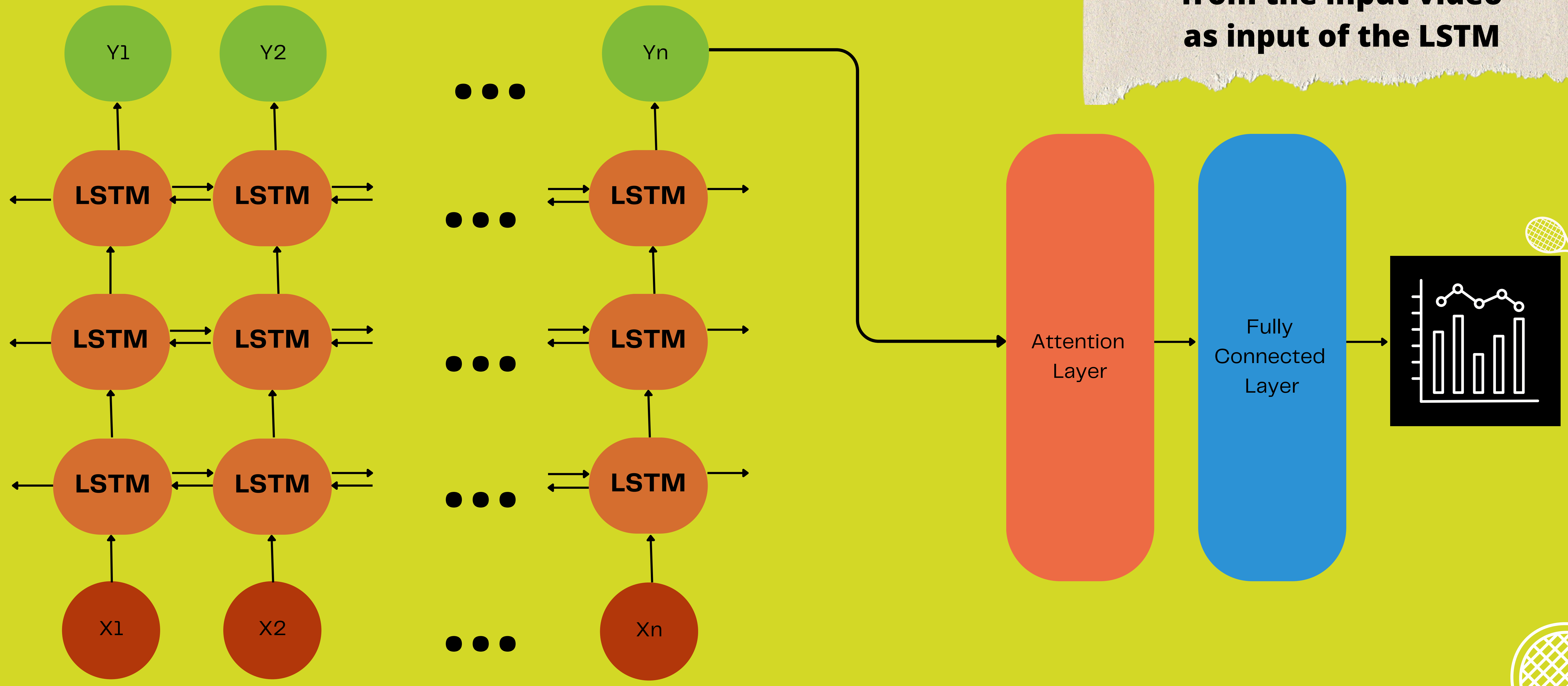
- **A sport based human action dataset composed of the 12 basic tennis shots captured by Kinect;**
- **The data are provided in 5 different synchronized forms (RGB, silhouettes, depth, 2D skeleton and 3D skeleton).**
- **Each shot has been performed several times resulting in 8734 videos (1980 RGB videos), converted to AVI format.**



12 types of shots:

- Backhand with two hands
- Backhand
- Backhand slice
- Backhand volley
- Forehand flat
- Forehand open stands
- Forehand slice
- Forehand volley
- Service flat
- Service kick
- Service slice
- Smash.

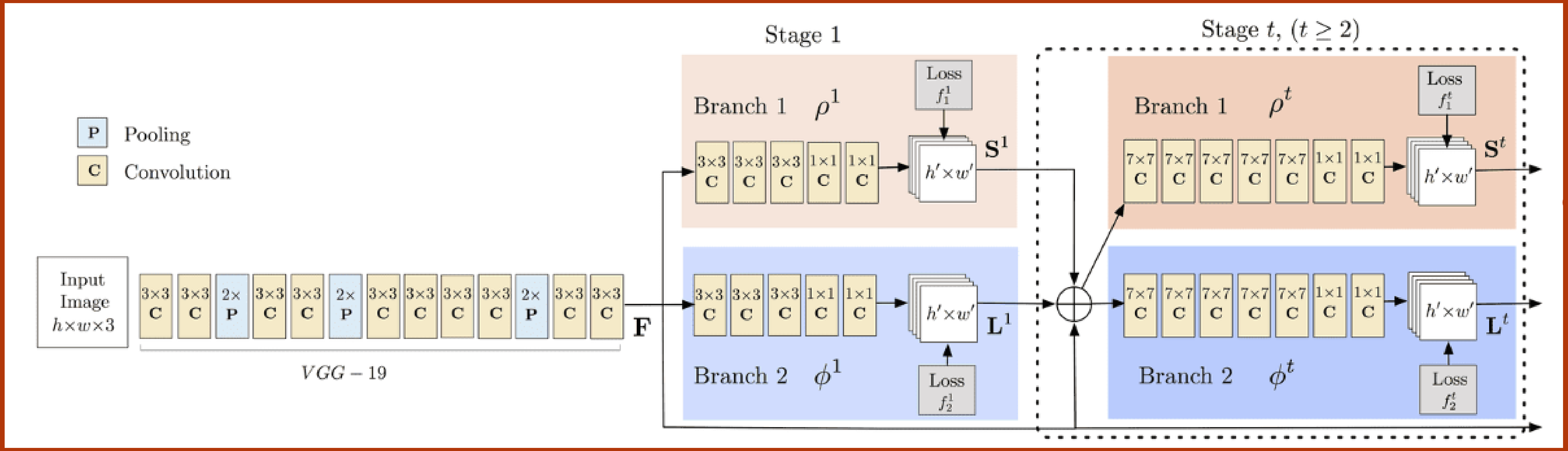
Architecture

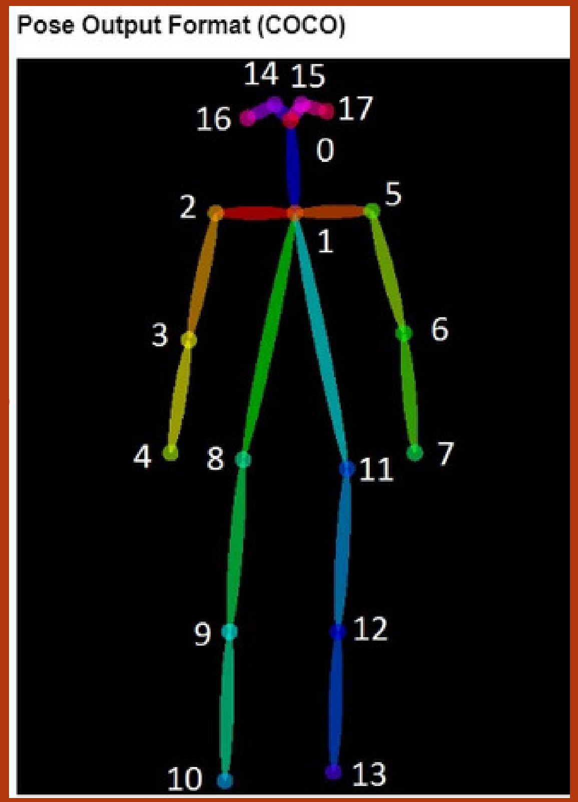


At each time step *i*, we have the frame feature vector *Xi* from the input video as input of the LSTM

OpenPose

- [2] Bottom-up approach for **multi-person human pose estimation**
- First detects the **keypoints** in the image
- Maps appropriate keypoints to **form pairs**
- Returns detected pose in **COCO pose output format**
- Each detected body joint has **(x, y, confidence_score)**





Preprocessing

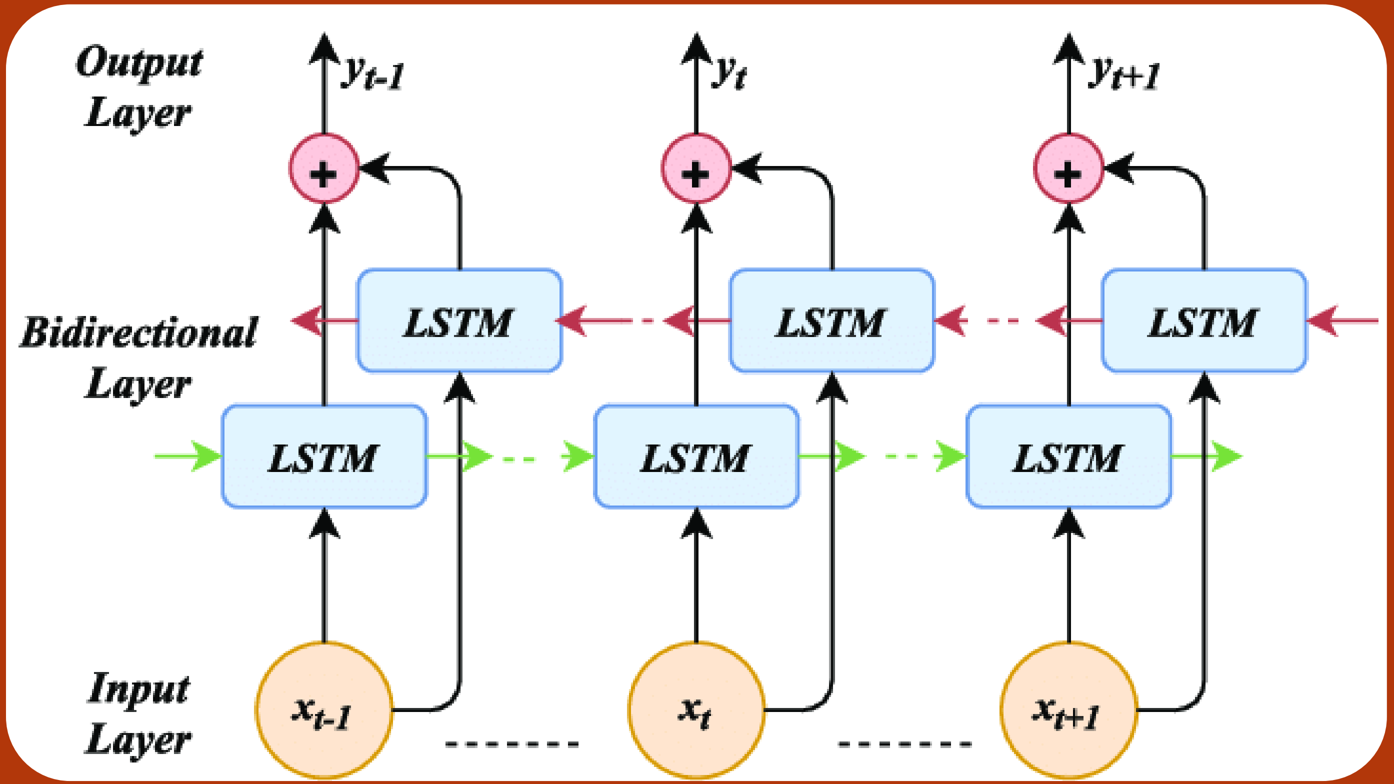
From OpenPose keypoints in json format to **dataset list**, composed by **1620 videos of 10 different types of shots**

Added **padding** where needed & made coordinates **invariant w.r.t. the position and size** of the tennis player's body.

Z-Score Normalization and **subdivision the dataset into train, validate** and **test** sets.

We don't consider the two classes Backhand slice Backhand volley

Bidirectional LSTM



with 3 layers

First: dropout2d & reshape

- Video as a sequence of N-frames, where a frame = 18 joints x 2 coordinates (x , y) , passed through a Dropout2D layer => some joints will be dropped out during training to prevent overfitting
- Reshape each frame from 18x2 to 36 coordinates.

Input:

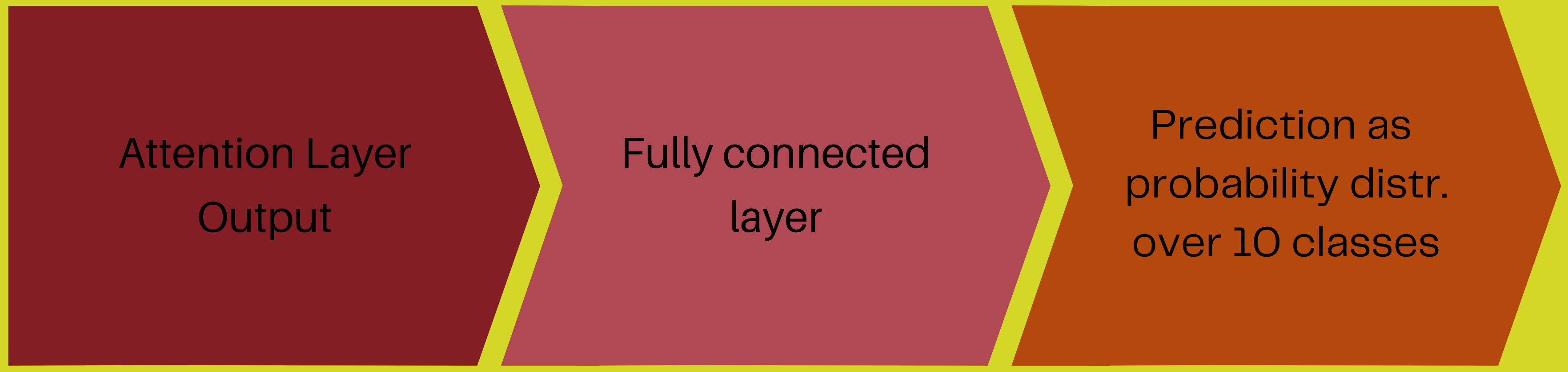
Video with size (N, 36) , in a PackedSequence object (to make padding frames irrelevant for the prediction)

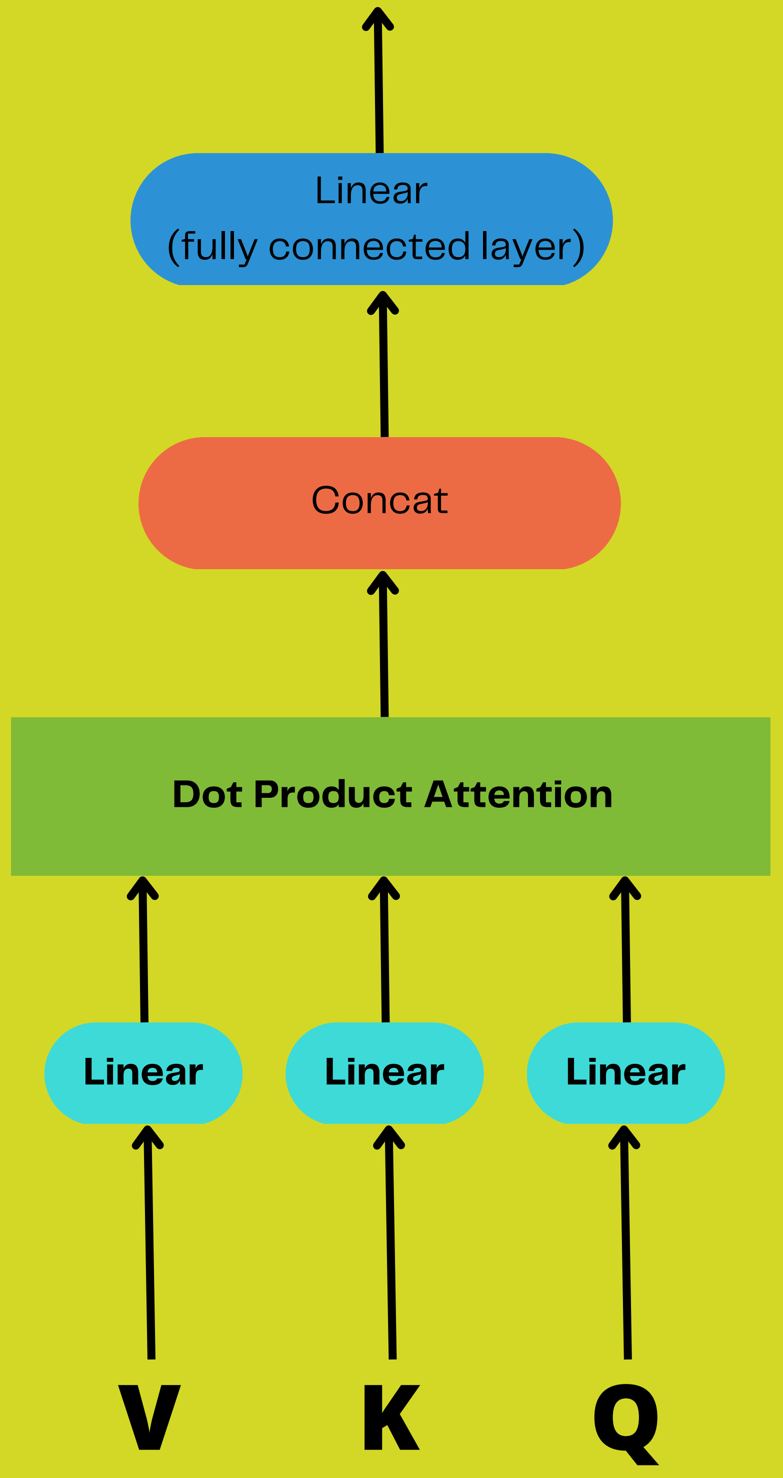
Output

1. Output features from the last layer, for each time-step t (corresponding to a frame)
2. Final hidden states (in both directions) of Bidirectional-LSTM, from the last of the three layers.

Attention Layer

- [3] Applied on the 2nd output of LSTM
- Query (Q) = 2nd output
- Key (K), Value (V) = 1st output
- Similarity measure: give more relevance to the output features that most influenced the final output





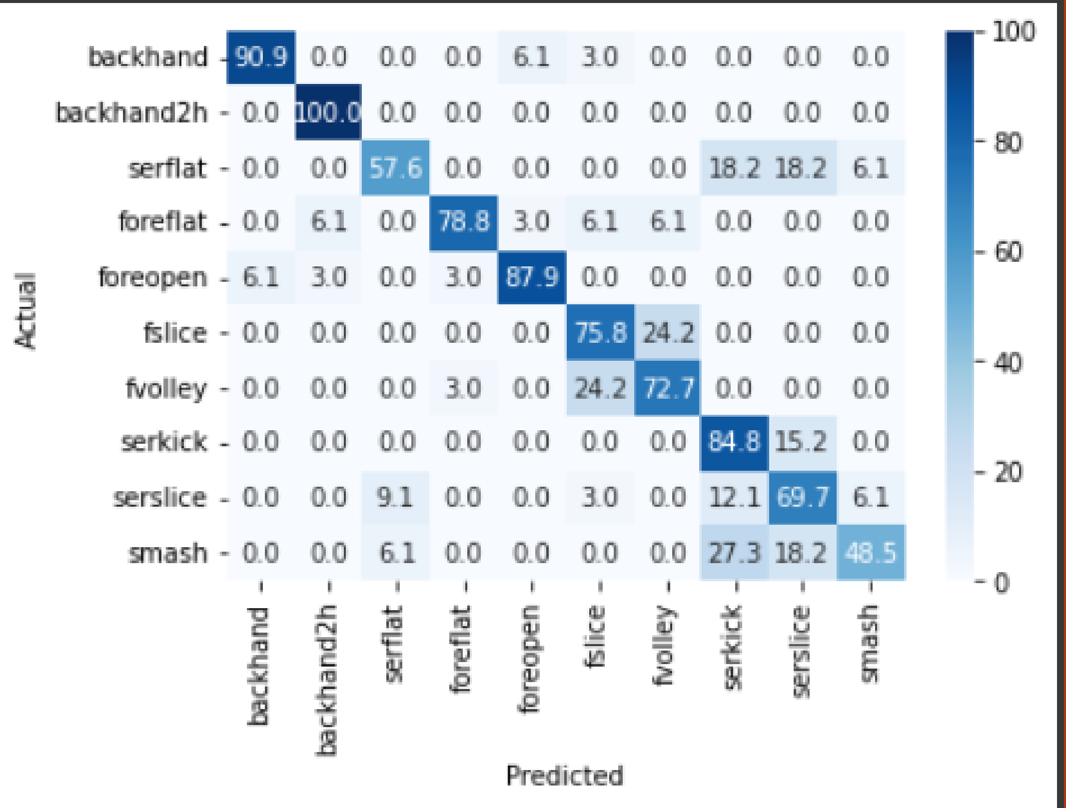
Model results and benchmarks

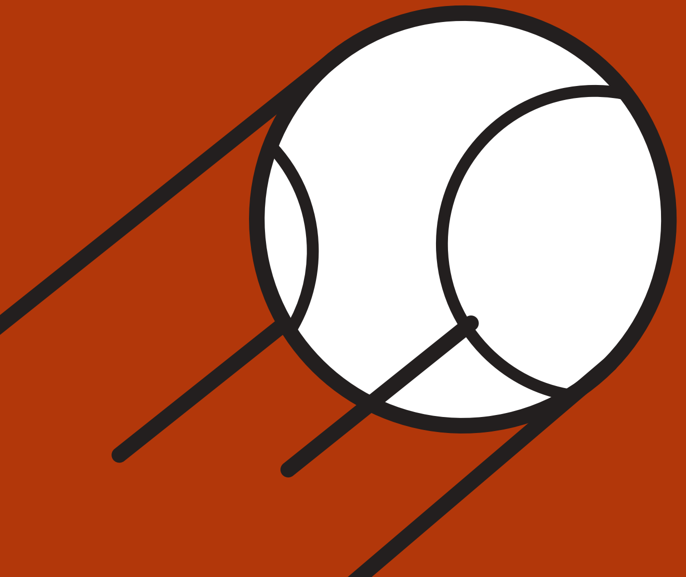
Starting from 1620 videos, we splitted the dataset in

- **48% training**
- **32% validating**
- **20% testing**

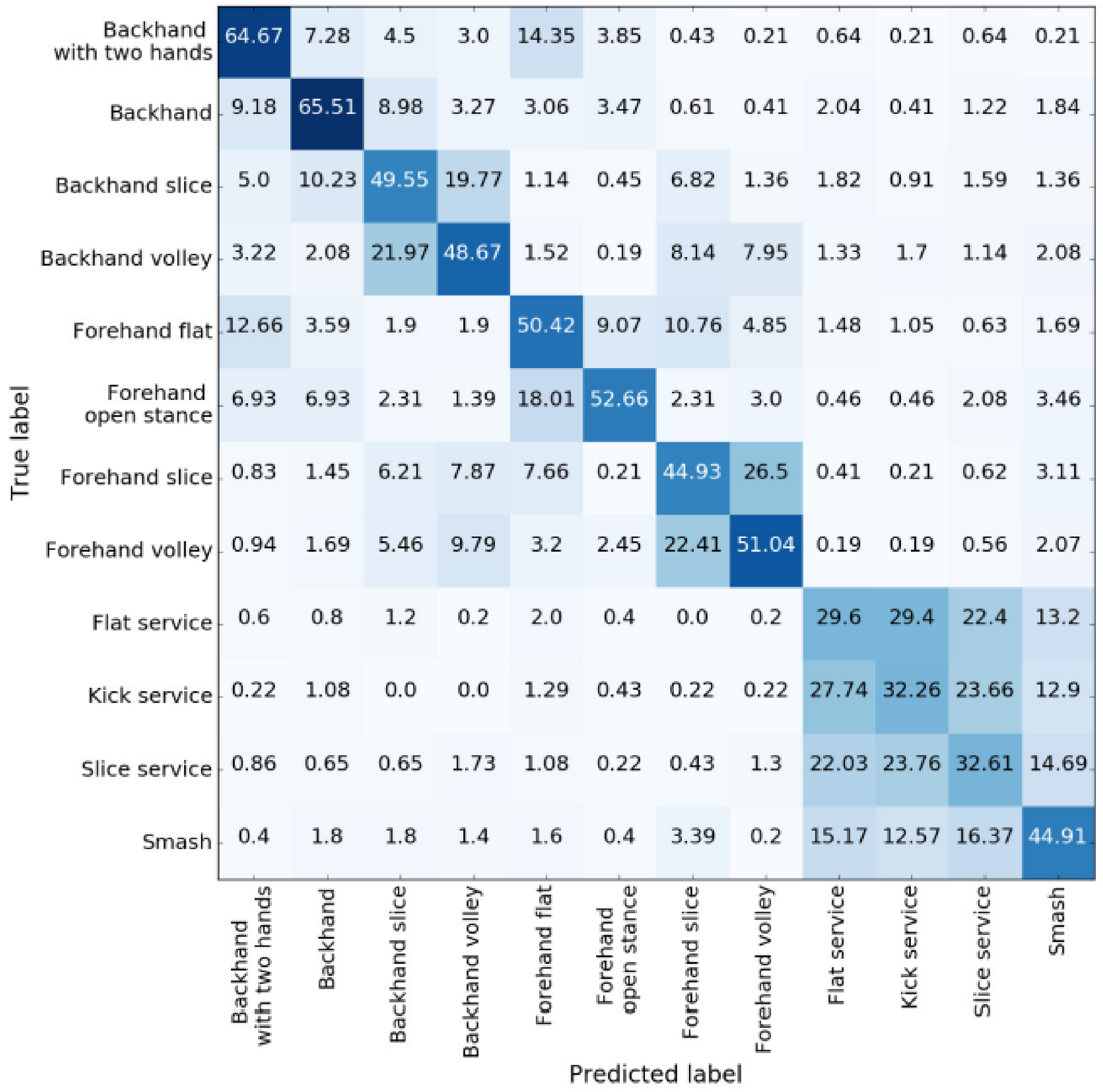
NUM_EPOCHS	300
NUM_HIDDEN	128
BATCH_SIZE	32
DROPOUT_2D	0.3
LSTM DROPOUT	0.5
Optimizer	Adagrad
LEARNING_RATE	0.001

```
test Loss: 0.7056 Accuracy: 0.7623 F1_Score: 0.6726
+------------+-----------+--------+
|    Shot    | Precision | Recall |
+------------+-----------+--------+
|  backhand  |   0.94    |  0.91  |
| backhand2h |   0.9     |  1.0   |
|  serflat   |   0.79    |  0.58  |
|  foreflat  |   0.93    |  0.79  |
|  foreopen  |   0.91    |  0.88  |
|   fslice   |   0.68    |  0.76  |
|  fvolley   |   0.71    |  0.73  |
|  serkick   |   0.6     |  0.85  |
|  serslice  |   0.57    |  0.7   |
|   smash    |   0.8     |  0.48  |
+------------+-----------+--------+
Confusion Matrix:
```





Deep Learning for Domain-Specific Action Recognition in Tennis [4]



	Backhand with two hands	Backhand	Backhand slice	Backhand volley	Forehand flat	Forehand open stance	Forehand slice	Forehand volley	Flat service	Kick service	Slice service	Smash
Backhand with two hands	64.67	7.28	4.5	3.0	14.35	3.85	0.43	0.21	0.64	0.21	0.64	0.21
Backhand	9.18	65.51	8.98	3.27	3.06	3.47	0.61	0.41	2.04	0.41	1.22	1.84
Backhand slice	5.0	10.23	49.55	19.77	1.14	0.45	6.82	1.36	1.82	0.91	1.59	1.36
Backhand volley	3.22	2.08	21.97	48.67	1.52	0.19	8.14	7.95	1.33	1.7	1.14	2.08
Forehand flat	12.66	3.59	1.9	1.9	50.42	9.07	10.76	4.85	1.48	1.05	0.63	1.69
Forehand open stance	6.93	6.93	2.31	1.39	18.01	52.66	2.31	3.0	0.46	0.46	2.08	3.46
Forehand slice	0.83	1.45	6.21	7.87	7.66	0.21	44.93	26.5	0.41	0.21	0.62	3.11
Forehand volley	0.94	1.69	5.46	9.79	3.2	2.45	22.41	51.04	0.19	0.19	0.56	2.07
Flat service	0.6	0.8	1.2	0.2	2.0	0.4	0.0	0.2	29.6	29.4	22.4	13.2
Kick service	0.22	1.08	0.0	0.0	1.29	0.43	0.22	0.22	27.74	32.26	23.66	12.9
Slice service	0.86	0.65	0.65	1.73	1.08	0.22	0.43	1.3	22.03	23.76	32.61	14.69
Smash	0.4	1.8	1.8	1.4	1.6	0.4	3.39	0.2	15.17	12.57	16.37	44.91

Figure 5. Confusion matrix of our model applied to the THETIS dataset.

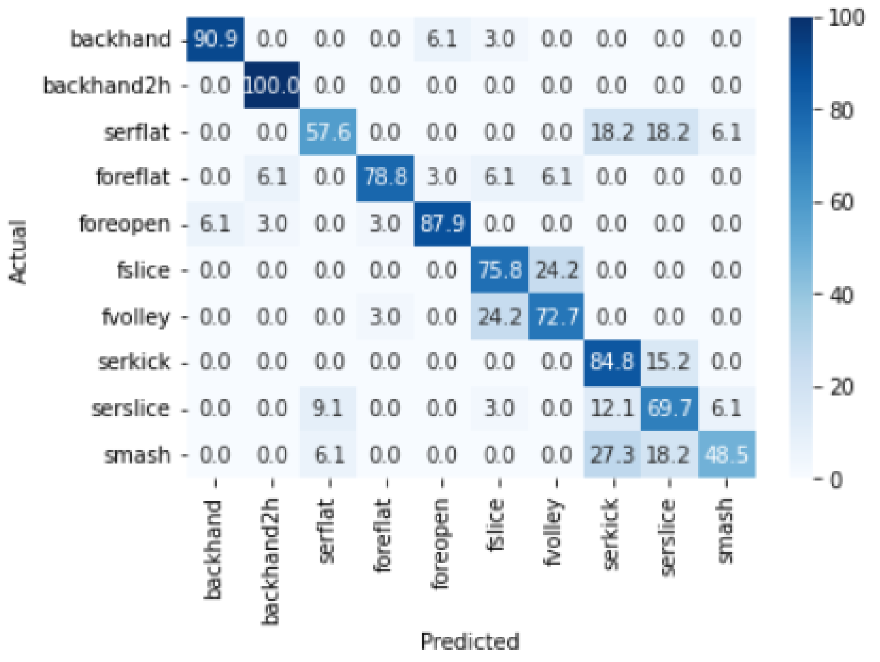
- **Always 3-layered LSTM for classification**
- **"Inception Neural Network" used for feature extraction**
- **All 12 shot classes analyzed**
- **47% Average Accuracy**

Our model

```
test Loss: 0.7056 Accuracy: 0.7623 F1_Score: 0.6726
```

Shot	Precision	Recall
backhand	0.94	0.91
backhand2h	0.9	1.0
serflat	0.79	0.58
foreflat	0.93	0.79
foreopen	0.91	0.88
fslice	0.68	0.76
fvolley	0.71	0.73
serkick	0.6	0.85
serslice	0.57	0.7
smash	0.8	0.48

Confusion Matrix:



	backhand	backhand2h	serflat	foreflat	foreopen	fslice	fvolley	serkick	serslice	smash
backhand	90.9	0.0	0.0	0.0	6.1	3.0	0.0	0.0	0.0	0.0
backhand2h	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
serflat	0.0	0.0	57.6	0.0	0.0	0.0	0.0	18.2	18.2	6.1
foreflat	0.0	6.1	0.0	78.8	3.0	6.1	6.1	0.0	0.0	0.0
foreopen	6.1	3.0	0.0	3.0	87.9	0.0	0.0	0.0	0.0	0.0
fslice	0.0	0.0	0.0	0.0	0.0	75.8	24.2	0.0	0.0	0.0
fvolley	0.0	0.0	0.0	3.0	0.0	24.2	72.7	0.0	0.0	0.0
serkick	0.0	0.0	0.0	0.0	0.0	0.0	0.0	84.8	15.2	0.0
serslice	0.0	0.0	9.1	0.0	0.0	3.0	0.0	12.1	69.7	6.1
smash	0.0	0.0	6.1	0.0	0.0	0.0	0.0	27.3	18.2	48.5

RGB Video Based Tennis Action Recognition Using a Deep Historical Long Short-Term Memory [5]

- **"Inception V3" model used for feature extraction**
- **"historical LSTM" model trained first with "HMDB51" dataset**
- **training on "THETIS" with 12 classes and K-Fold Cross Validation**

TABLE II
ACCURACY COMPARISON OF METHODS ON THETIS DATASET

Method	Accuracy
Historical LSTM ($\tau = 2$)	0.70
Historical LSTM ($\tau = 3$)	0.74
Historical LSTM ($\tau = 4$)	0.71
Historical LSTM ($\tau = 5$)	0.63
LSTM	0.56
Mora et al. [2]	0.47
Gourgari et al. (using depth videos)[8]	0.6
Gourgari et al. (using 3D skeletons)[8]	0.54

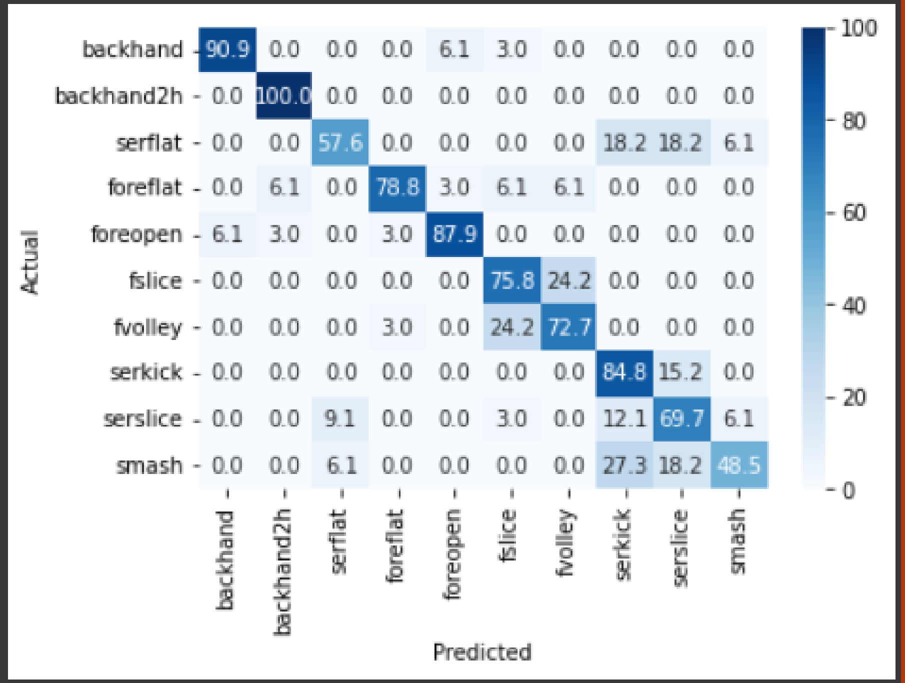
- **τ = length of state sequence used to re-initialize historical state, using the response states from time t − τ to time t**

Our model

test Loss: 0.7056 Accuracy: 0.7623 F1_Score: 0.6726

Shot	Precision	Recall
backhand	0.94	0.91
backhand2h	0.9	1.0
serflat	0.79	0.58
foreflat	0.93	0.79
foreopen	0.91	0.88
fslice	0.68	0.76
fvolley	0.71	0.73
serkick	0.6	0.85
serslice	0.57	0.7
smash	0.8	0.48

Confusion Matrix:



Thanks for your attention!

References

[1] THETIS

[2]OpenPose Documentation

[3]Dot Product Attention

[4]Deep Learning for Domain-Specific Action Recognition in Tennis

[5]RGB Video Based Tennis Action Recognition Using a Deep Historical Long Short-Term Memory