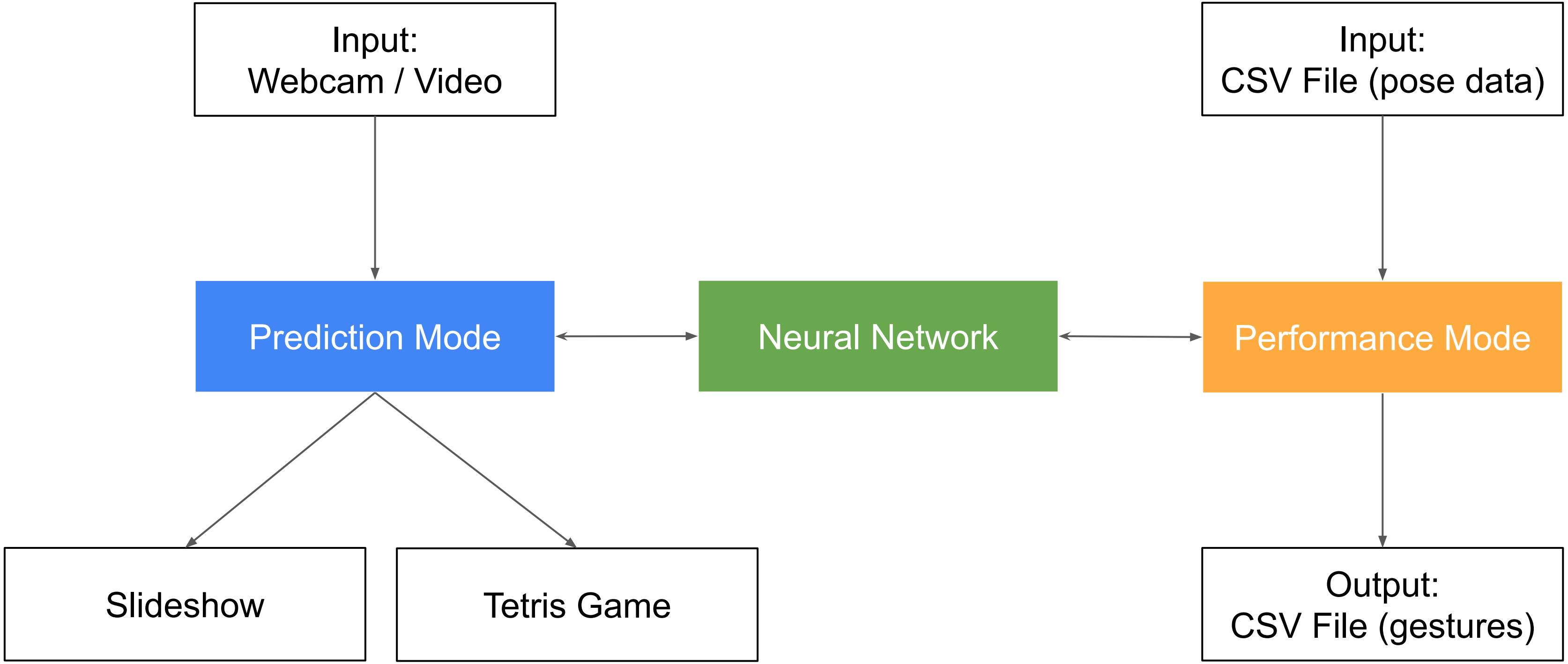
MLP Gesture Classifier

Controlling Interactive Applications

Marcel Roth, Micha Nowak, Jan-Philipp Friese

Project Overview



Input:
Webcam / Video

Prediction Mode

Neural Network

Performance Mode

Input:
CSV File (pose data)

Slideshow

Tetris Game

Output:
CSV File (gestures)

Data Acquisition

Gestures:

⟵ ⟶ ↻ ↺ ↑ ↓ ⤢ ⤡ ⟳ ☛

Raw Data Facts:

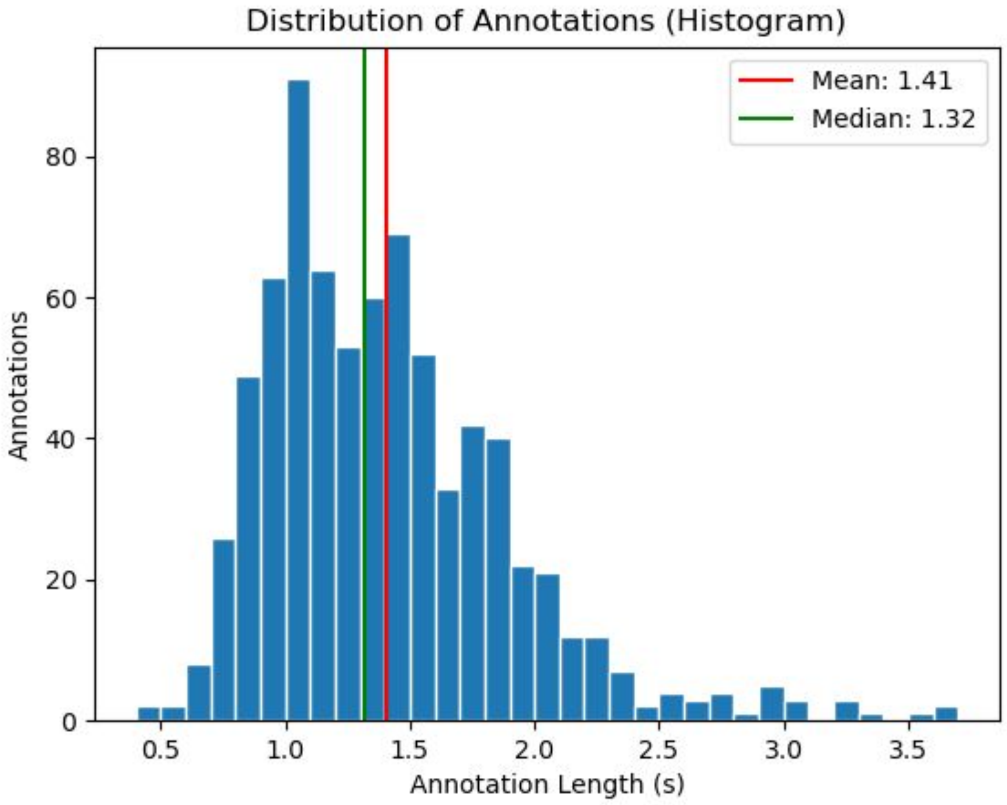
- 3 videos (each ~2mins) containing ~20 gestures each
- ~80 annotations per gesture (except point and spin)
- Converted videos to CSV files using **Mediapipe**

Annotation:

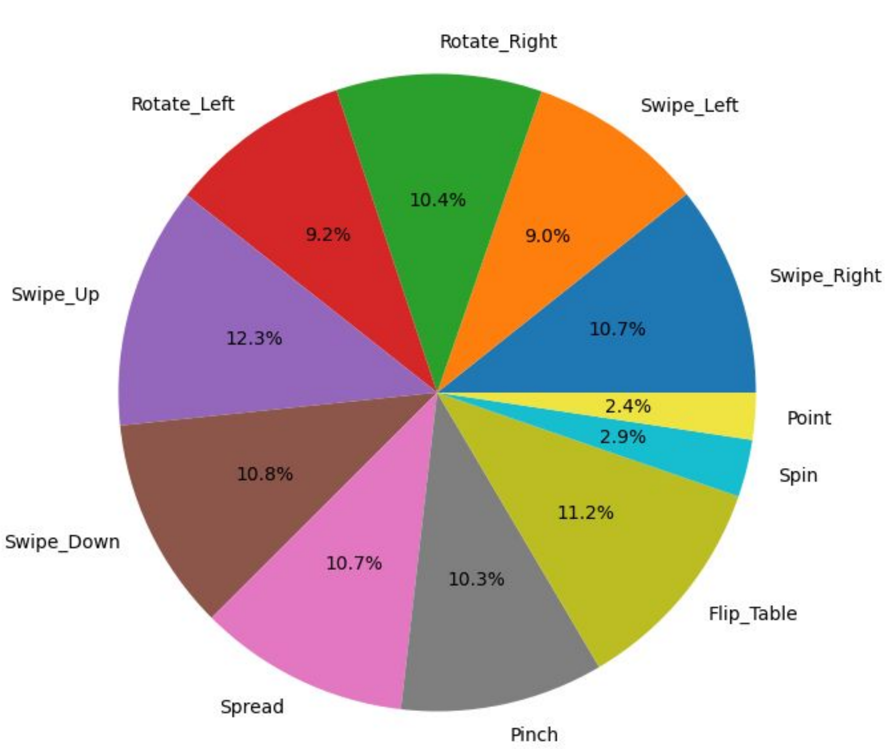
- Annotated gestures videos manually using **ELAN**

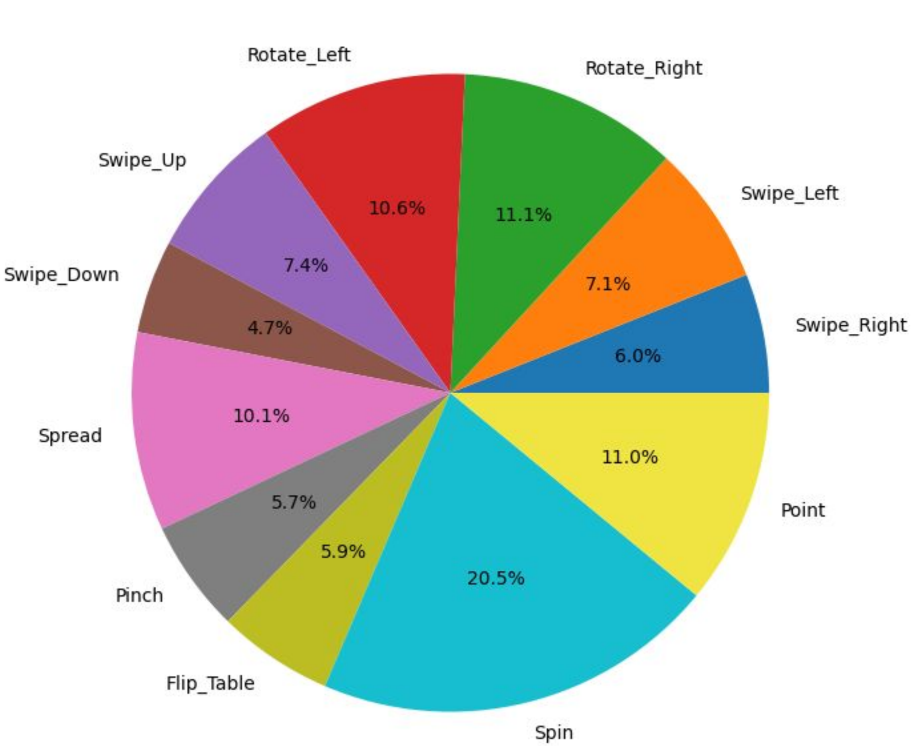
Data Acquisition

→ In Total **757** Annotations

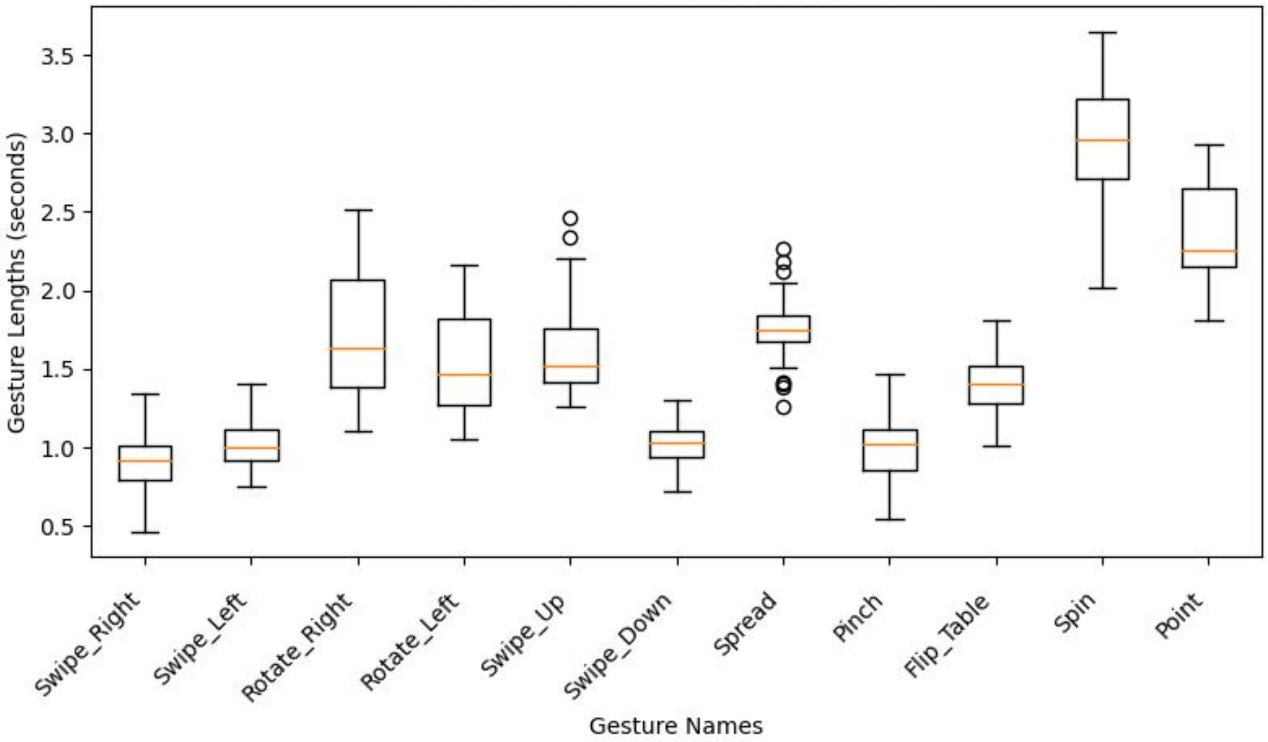


Data Acquisition: Distributions


Gesture Amounts

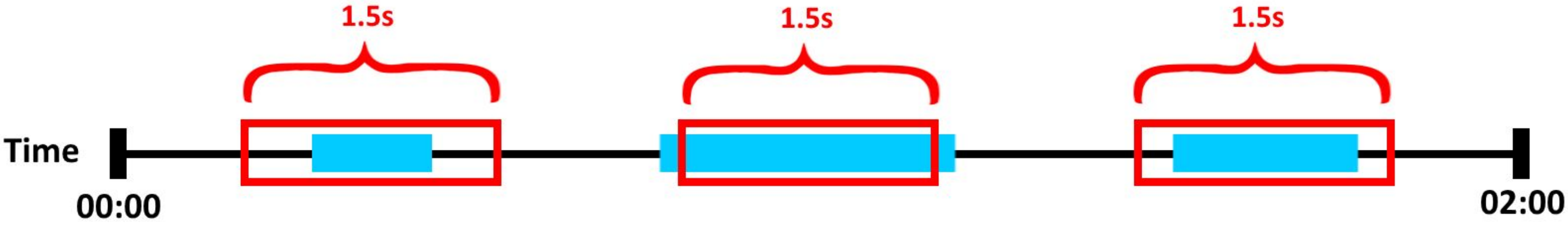

Average Gesture Annotation Lengths

Data Acquisition: Gesture Length Distribution



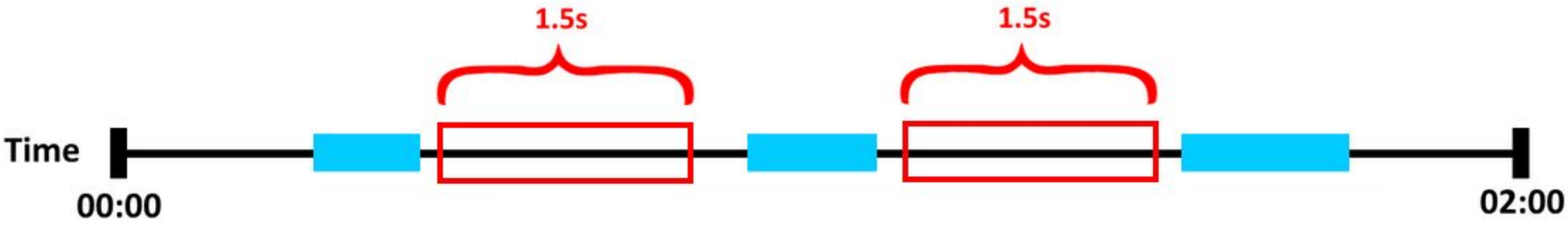
Data Preprocessing: Annotation → Sample

Positive Samples were extracted as 1.5s windows centered around the annotation



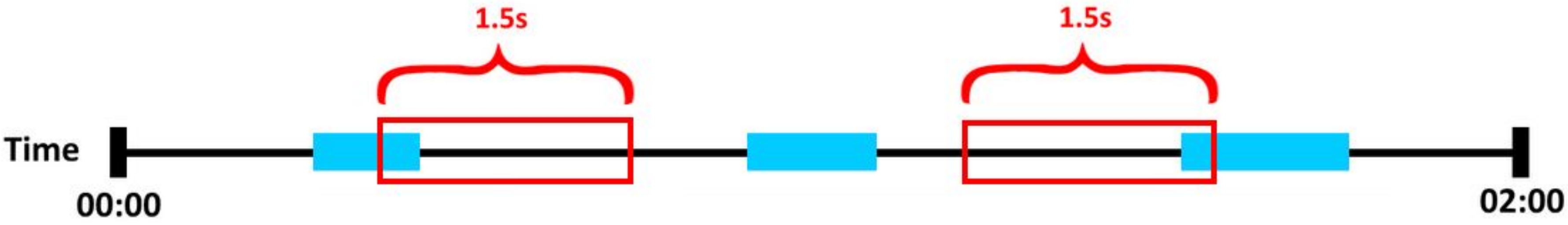
Data Preprocessing: Annotation → Sample

Idle Samples were extracted as 1.5s windows between annotations



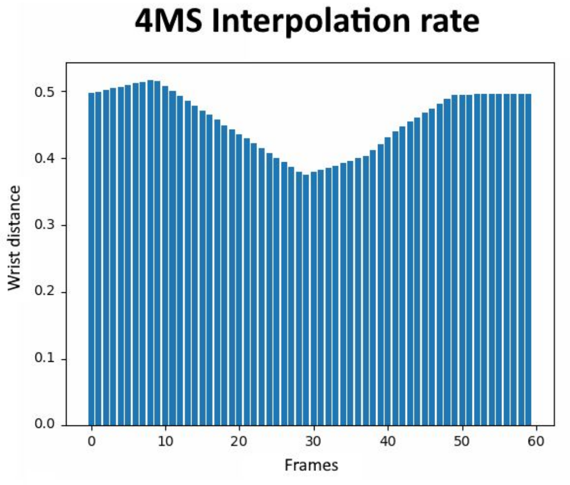
Data Preprocessing: Annotation → Sample

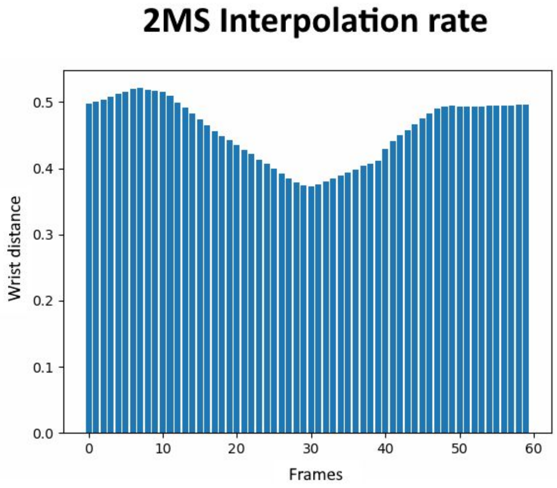
Overlapping Idle Samples were extracted as 1.5s windows, overlapping annotations by a random percentage (10-20%)

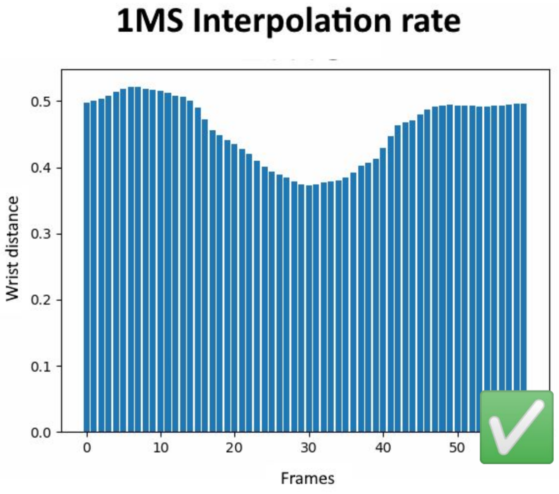


Data Preprocessing: Interpolation

Here: Value of the feature “wrist distance” for one sample, plotted over time

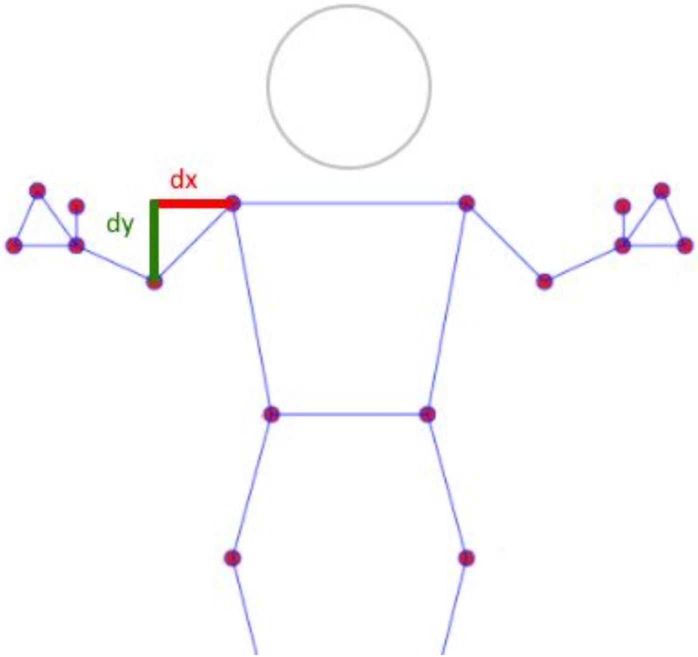

4MS Interpolation rate


2MS Interpolation rate


1MS Interpolation rate

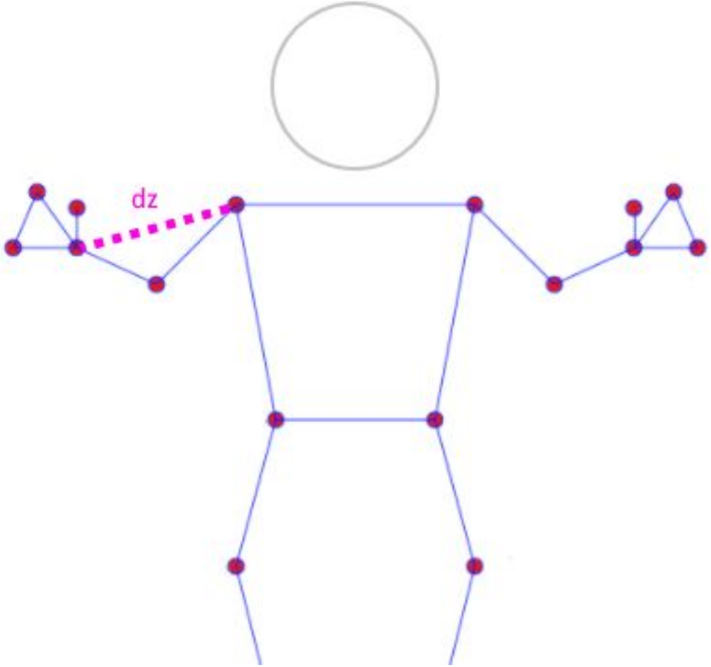
Data Preprocessing: Synthetic Features

X / Y differences for [Shoulder, Elbow], [Elbow Wrist], [Shoulder Wrist] (left and right)



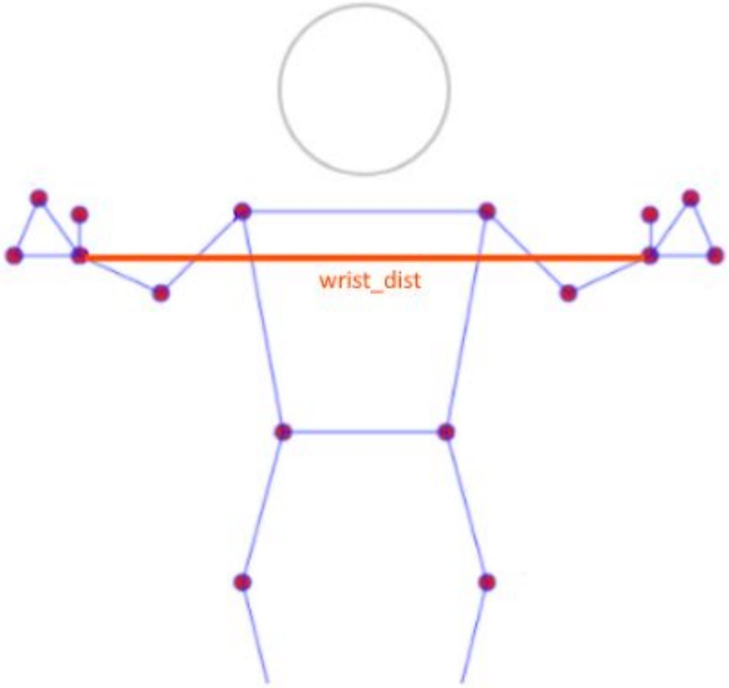
Data Preprocessing: Synthetic Features

Z differences for [Shoulder, Wrist] (left and right)



Data Preprocessing: Synthetic Features

Wrist distance



Data Preprocessing: Summary

- **Features per Frame:**
  - 12x (X, Y) differences [Shoulder, Elbow, Wrist]
  - 2x (Z) differences [Shoulder, Wrist]
  - 1x Wrist distance

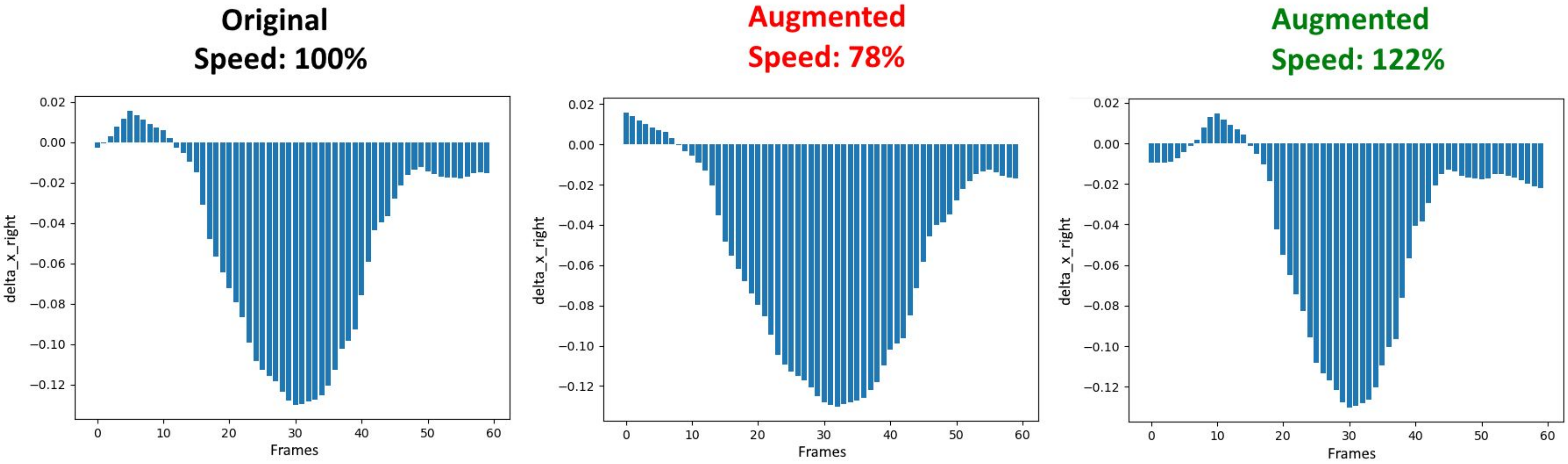
→ In Total 15 Float Values per Frame

<u>Assumption:</u> Window Size = 1.5 seconds
→ 60 Frames per Sample to achieve a “resolution” of **45 FPS**
→ 60 Frames per Sample x 15 Values per Frame

→ In Total 900 Features per Sample

Data Augmentation: Random Speed

Value of feature “delta_x_right” for one sample, plotted over time with different speeds



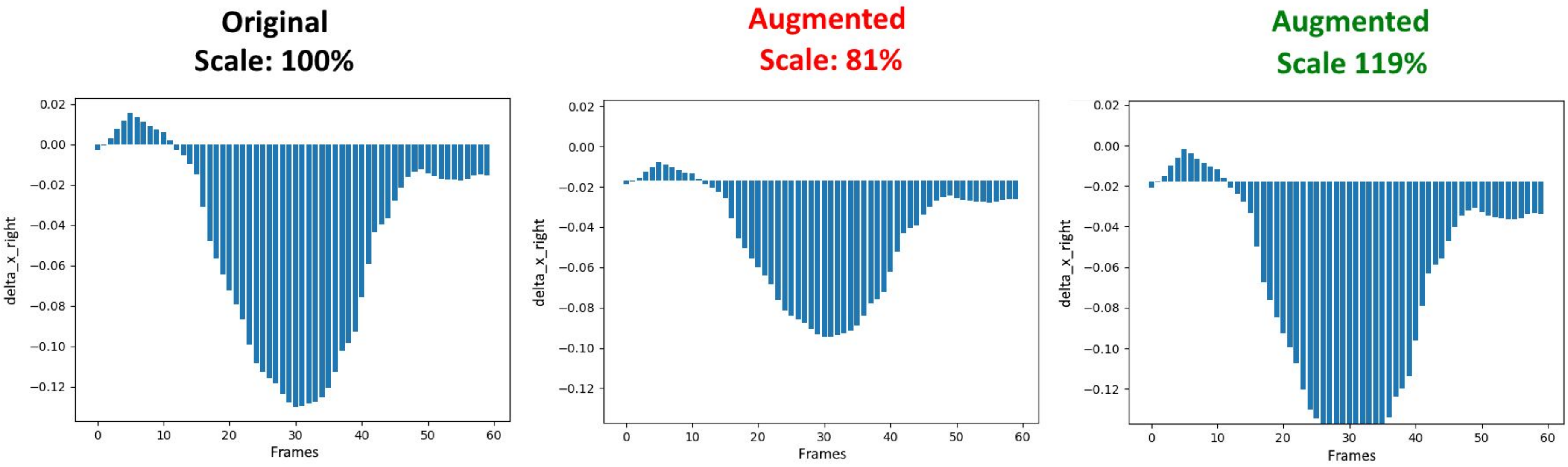
Original
Speed: 100%

Augmented
Speed: 78%

Augmented
Speed: 122%

Data Augmentation: Random Magnitude

Value of feature “delta_x_right” for one sample, plotted over time with different scales



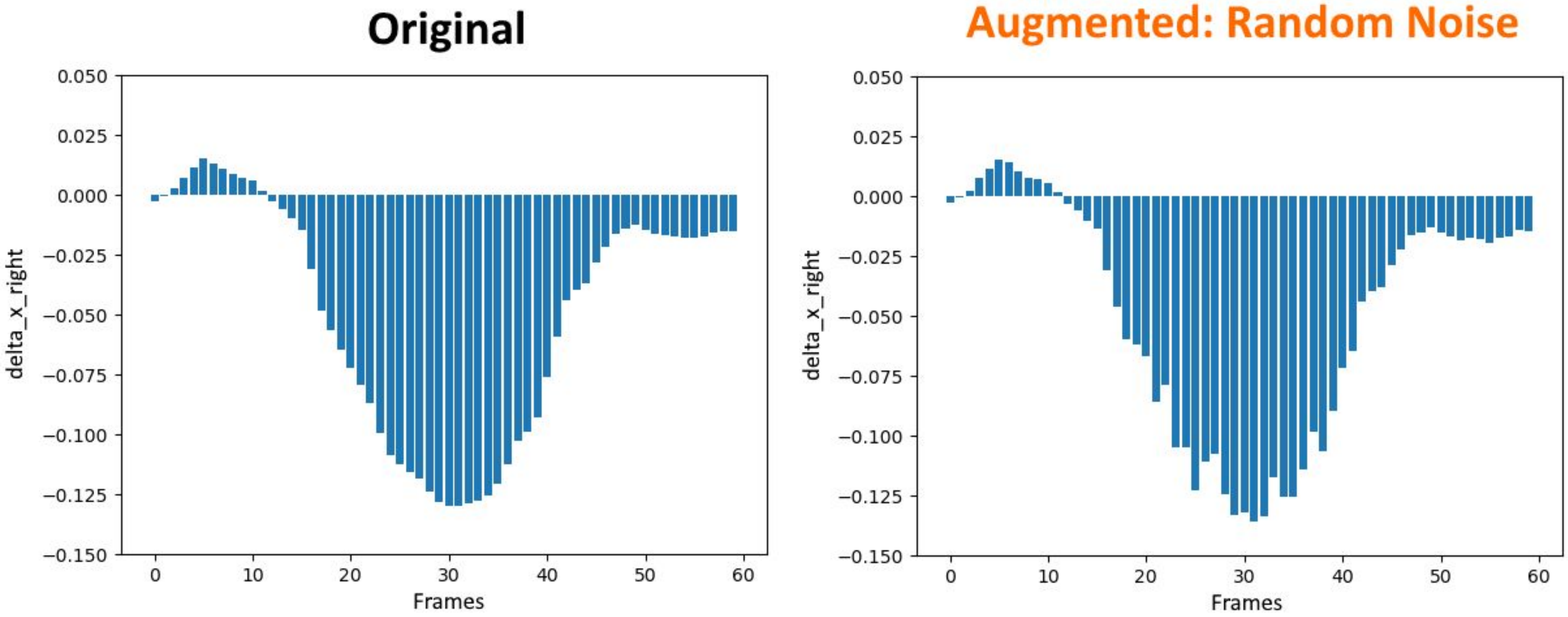
Original
Scale: 100%

Augmented
Scale: 81%

Augmented
Scale 119%

Data Augmentation: Random Noise

Value of feature “delta_x_right” for one sample, plotted over time without and with noise



Original

Augmented: Random Noise

FINAL DATASET

114936 SAMPLES

800MB

Training the Neural Network

- **Network Type**
    - Multilayer Perceptron (MLP)

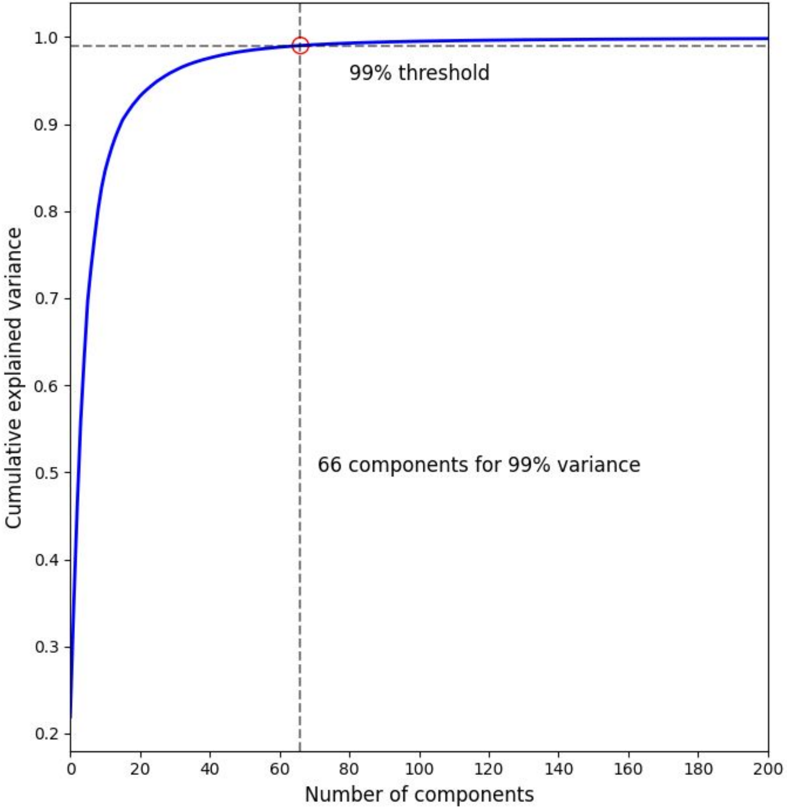
- **Activation Functions**
    - Hidden Layers → *Sigmoid*
    - Output Layer → *Softmax*

- **Loss Function**
    - Categorical Cross-Entropy

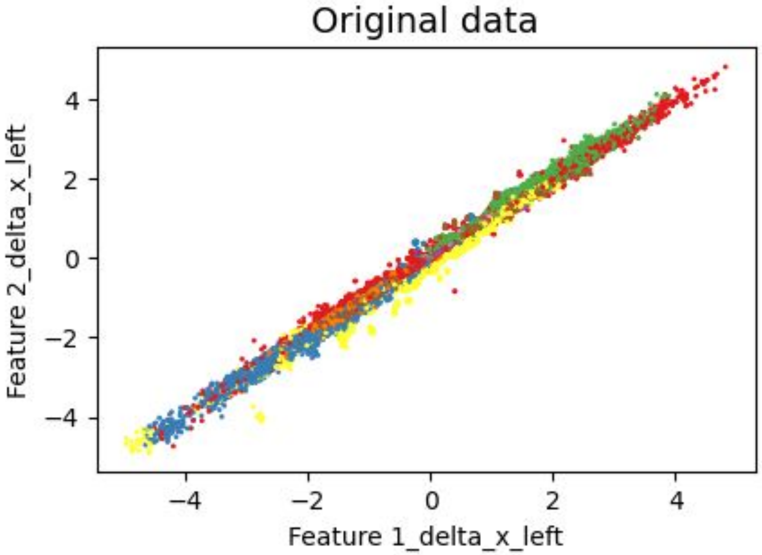
Principal Component Analysis

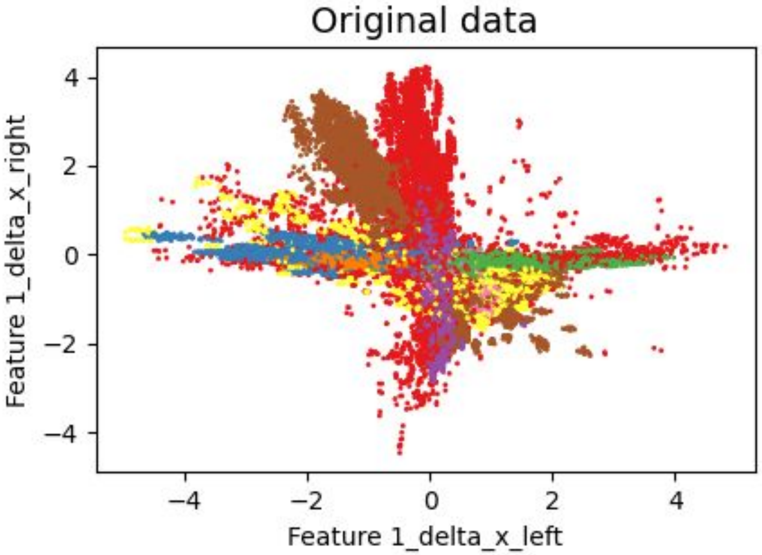
900 Features → **66** Principal Components

→ Capturing **99%** variance of the data

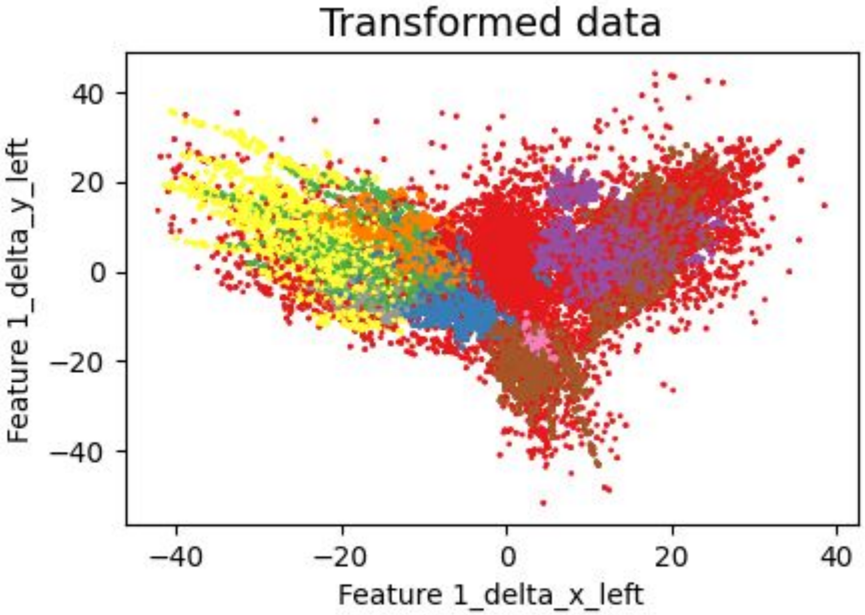


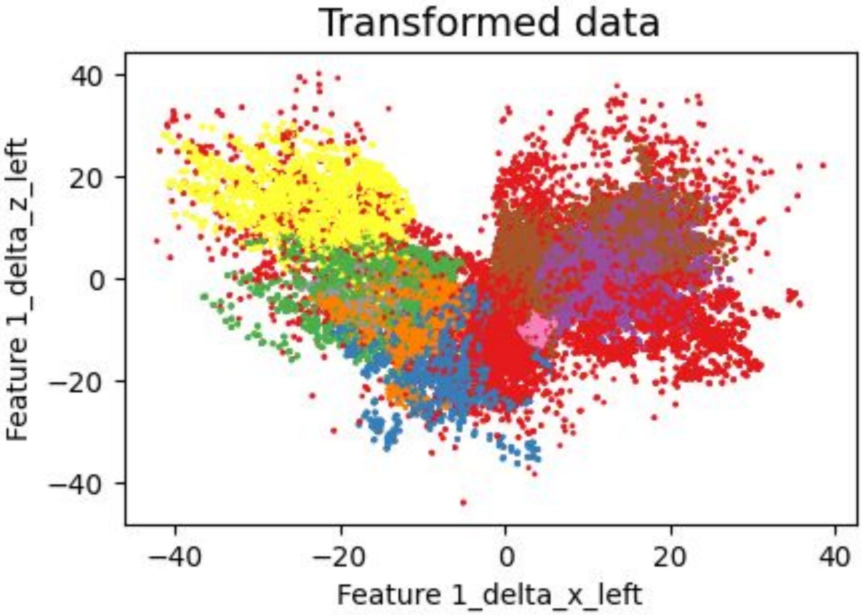
Principal Component Analysis



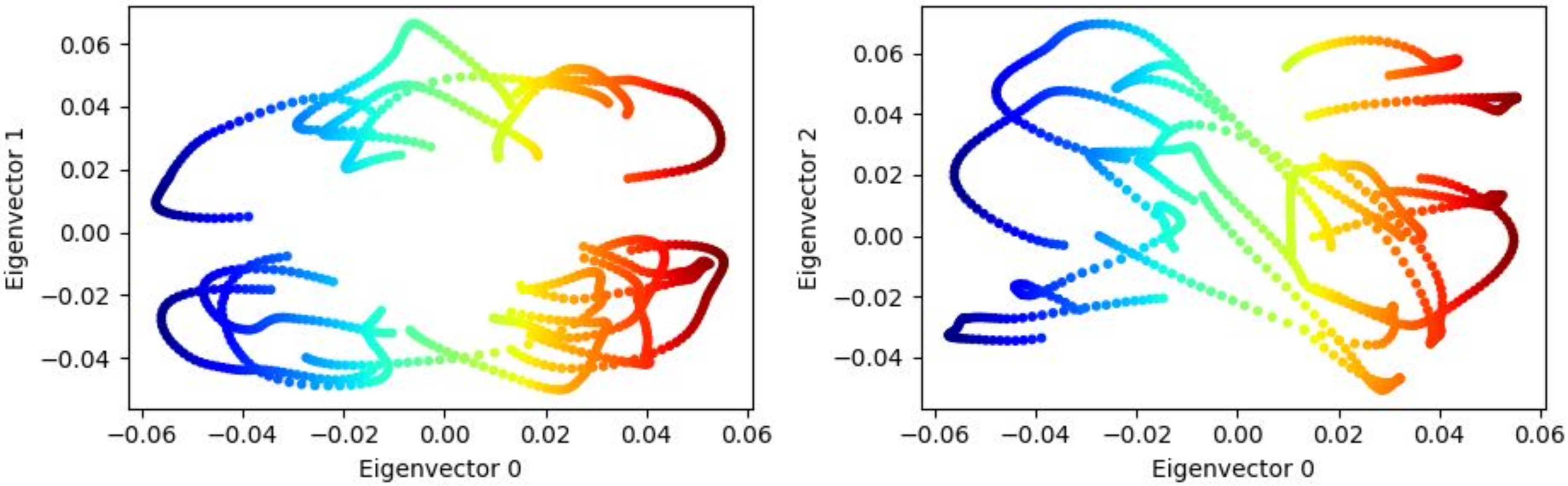


Principal Component Analysis





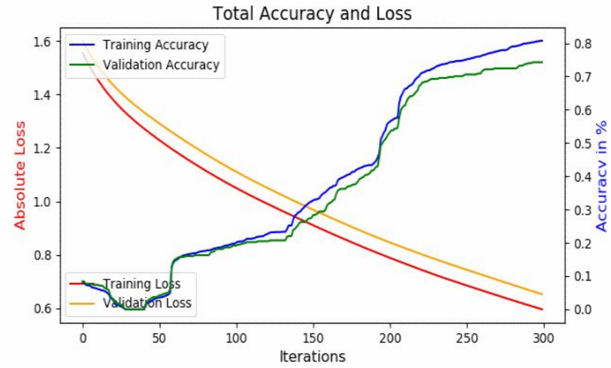
Principal Component Analysis: Eigenvectors



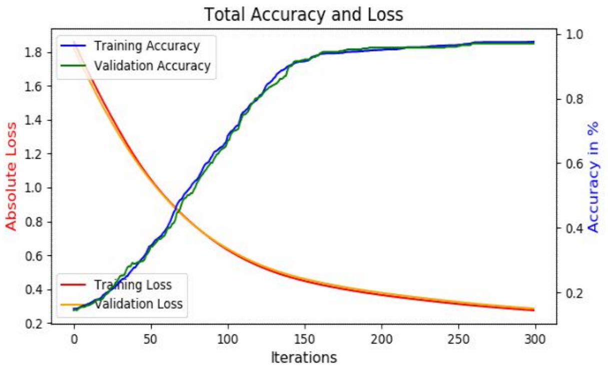
Hidden Layer Shape

Comparing varying hidden layer shapes with other hyperparameters fixed

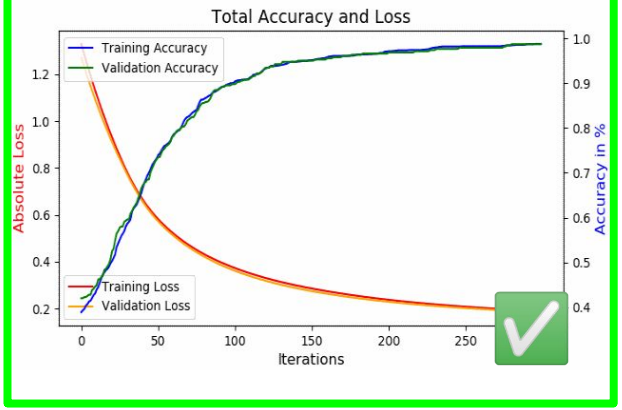
Hidden layer shape: 5



Hidden layer shape: 10



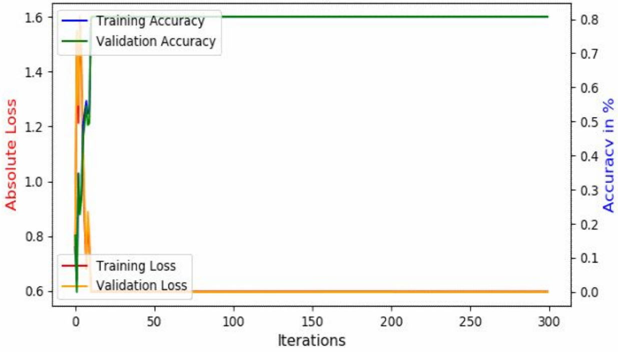
Hidden layer shape: 18



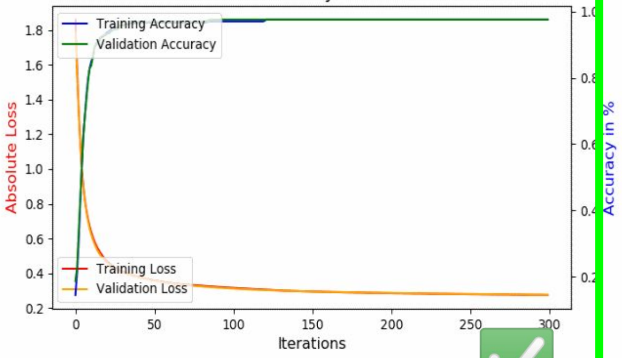
Learning Rate

Comparing varying learning rates with other hyperparameters fixed

Learning rate: 0.01

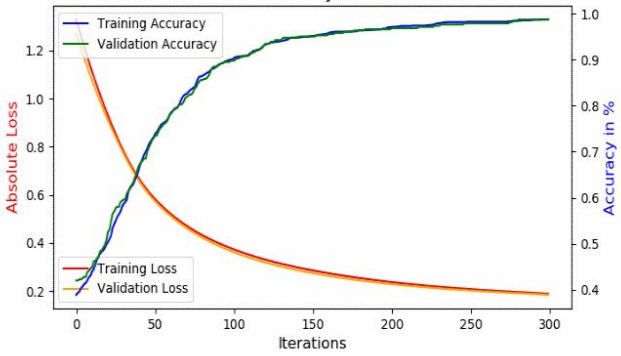


Learning rate: 0.001



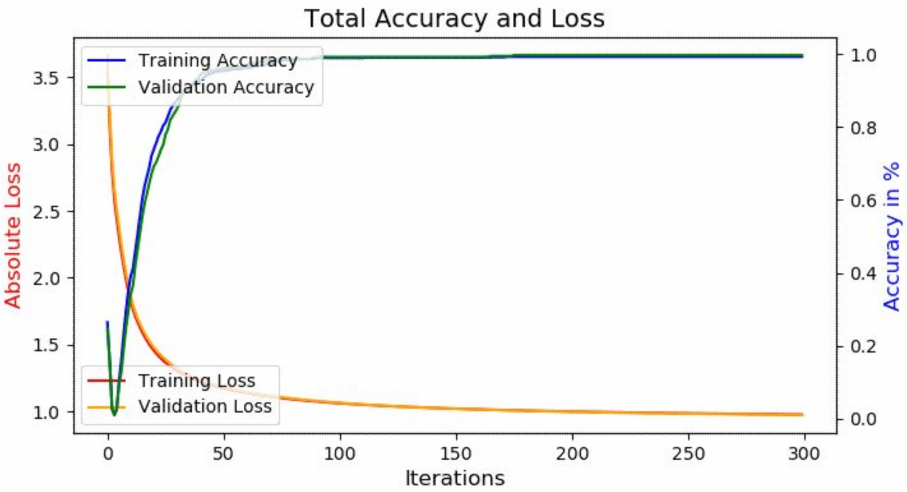
✅

Learning rate: 0.0001

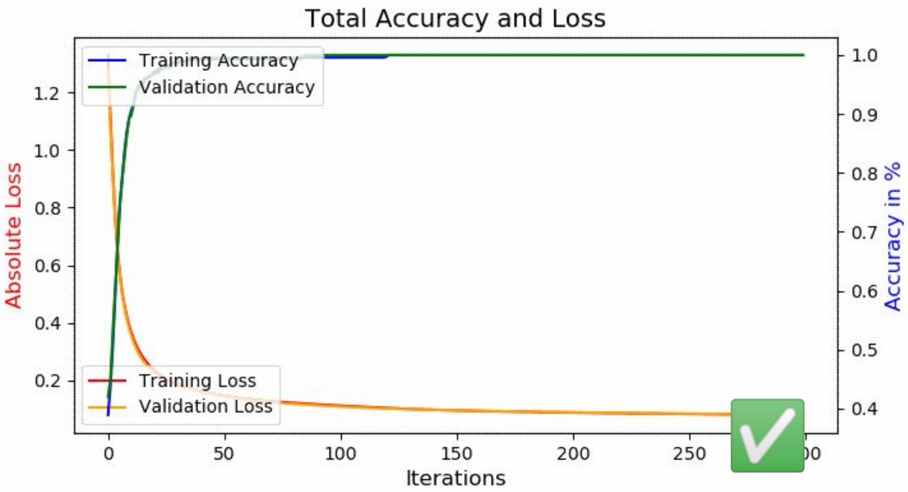


Feature Scaling

Without Feature Scaling

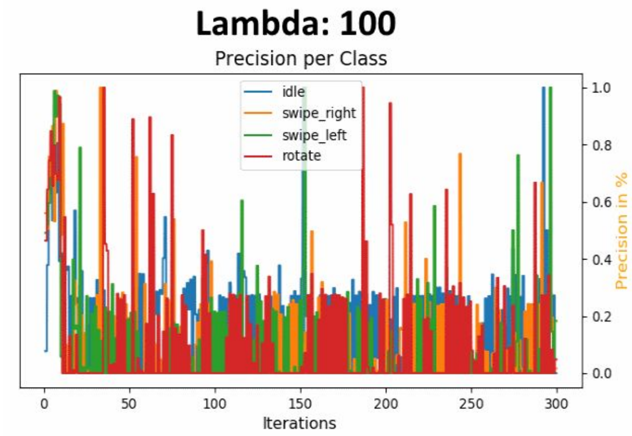


With Feature Scaling

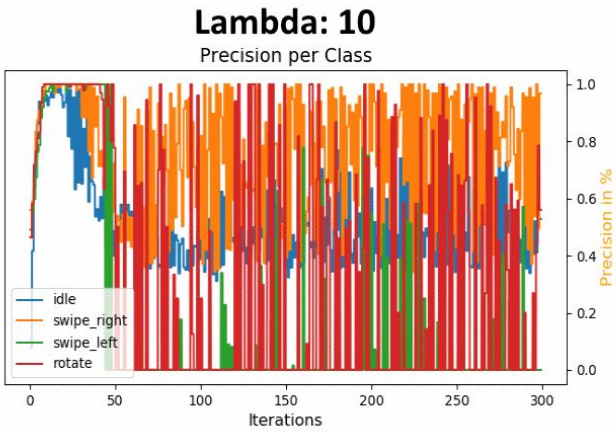


Regularization (L1)

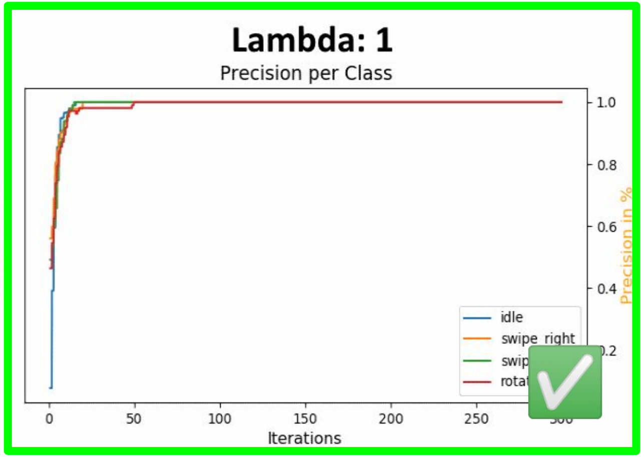
Comparing varying lambda values with other hyperparameters fixed


Lambda: 100

Precision per Class


Lambda: 10

Precision per Class


Lambda: 1

Precision per Class

✅

Overview: Dataset Creation Hyperparameters

Window size

Frames per sample

Interpolation rate

Positive shifted samples amount

Positive shift max %

Positive speed variations

Positive speed max %

Positive scaled samples

Positive scale max %

Positive noisy samples

Positive noise max %

Idle samples per file

Overlapping idle samples

Overlap max %

Overlap min %

Static idles per file

Idle scaled samples

Idle scale max %

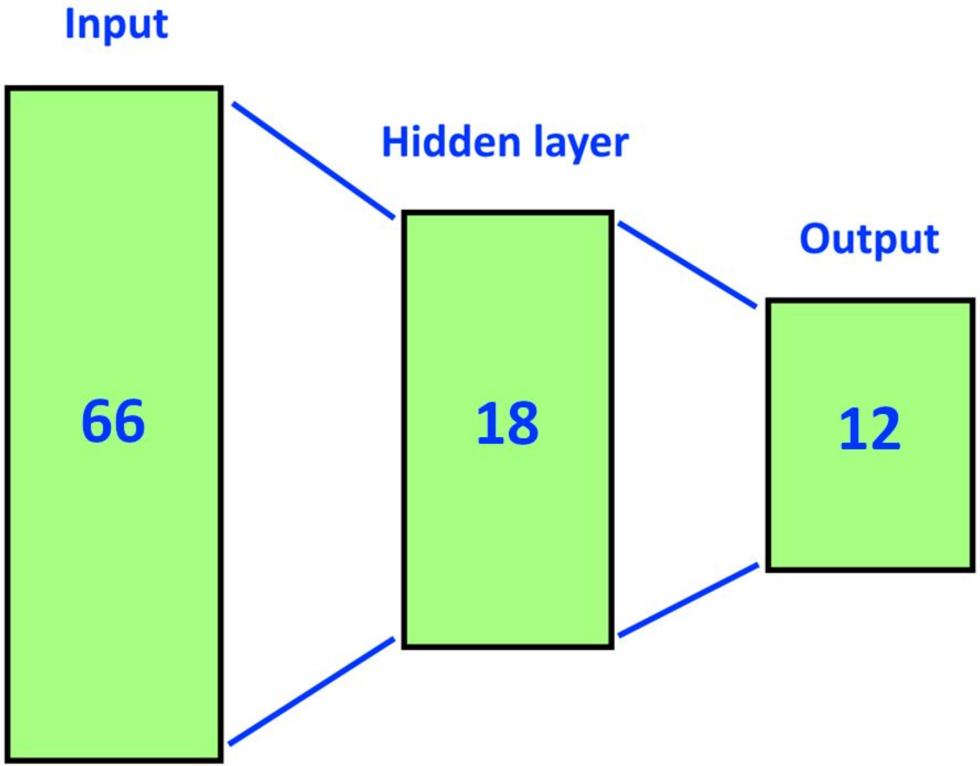
Idle noisy samples

Idle noise max %

Final Model: Training Hyperparameters

Network Seed	=	12
Hidden Layer Shape	=	[18]
Iterations	=	2000
Alpha	=	0.00003
Feature Scaling	=	True
Regularization	=	True
Lambda	=	0.0001
PCA Threshold	=	0.99
Validation Split	=	0.2

Final Model: Network Shape

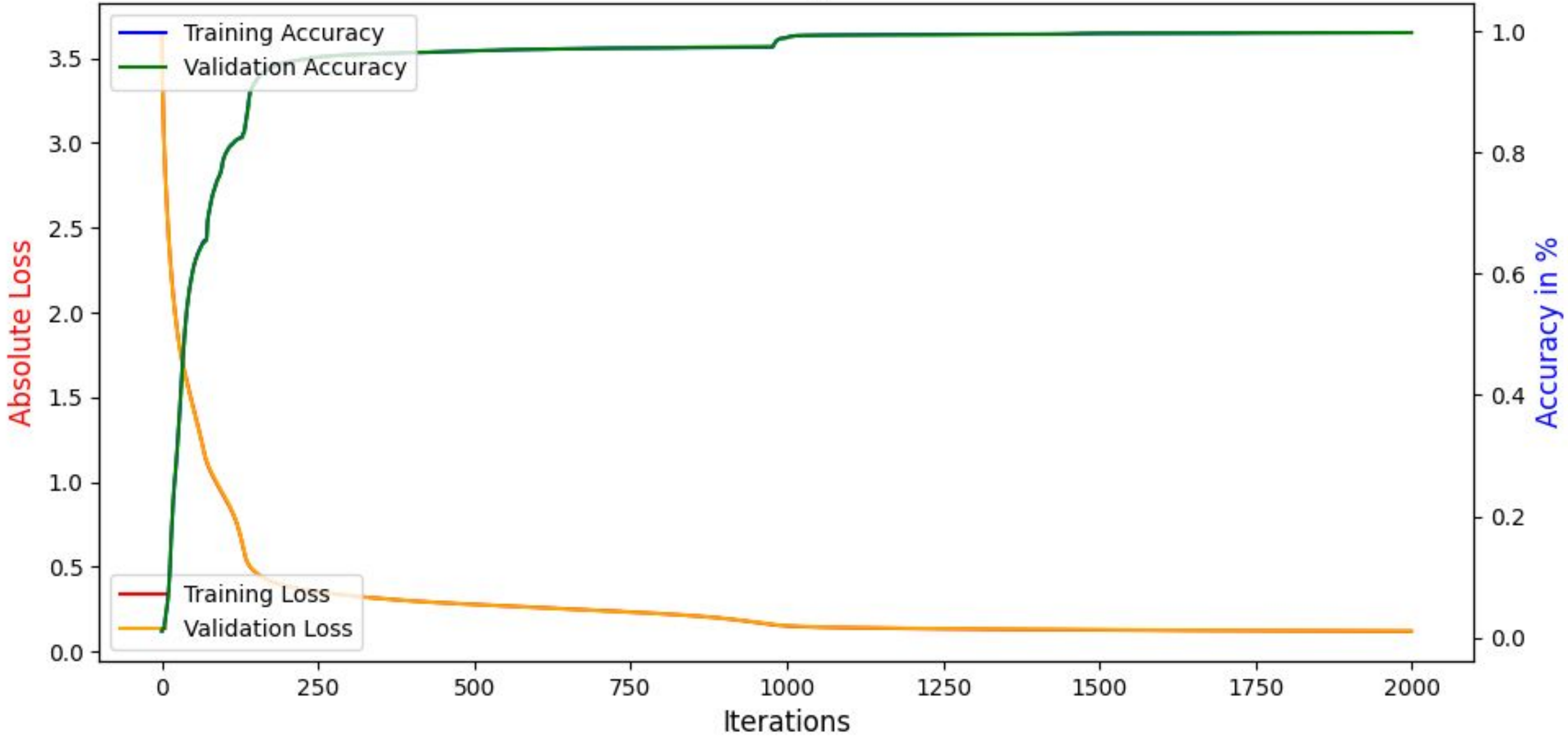


Final Model: Metrics

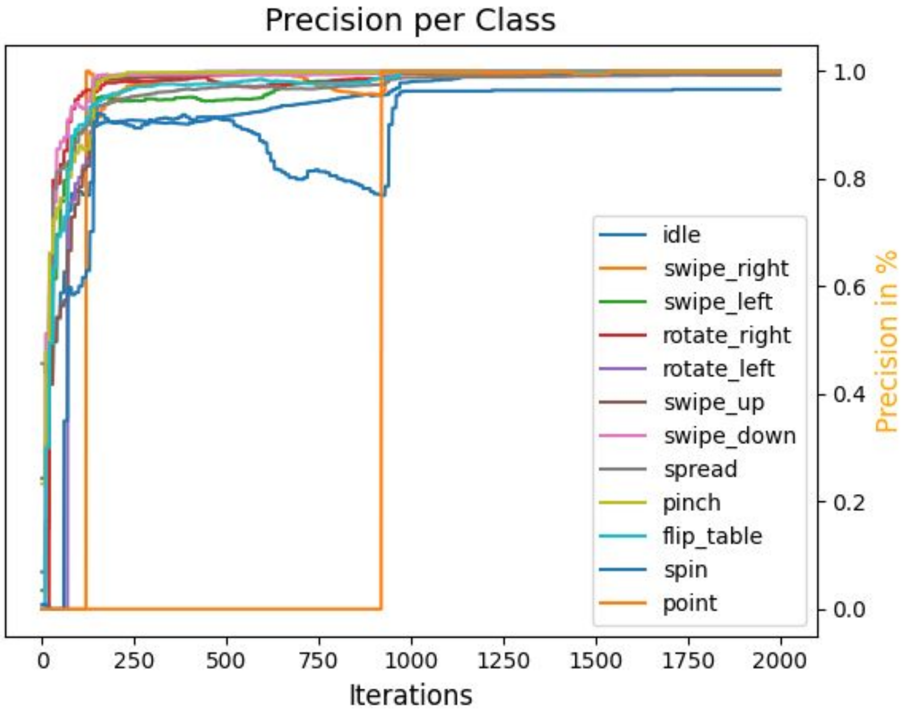
Final Training Accuracy:	99.74%
Final Validation Accuracy:	99.73%

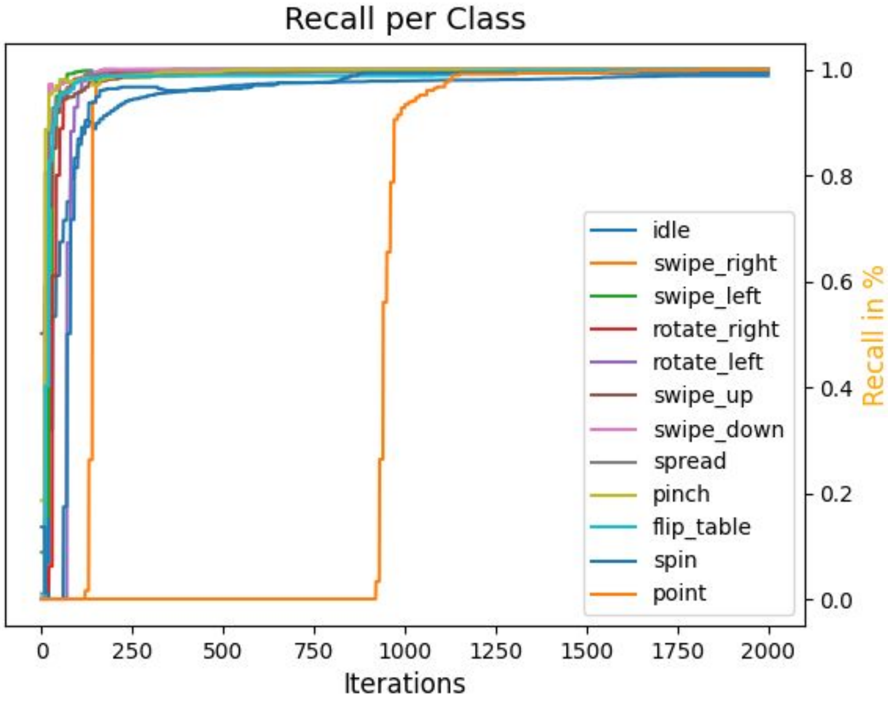
Final Recall Per-Class:	99.82%
Final Precision Per-Class:	99.62%
Final F1 Score Per-Class:	99.71%

Training: Loss and Accuracy

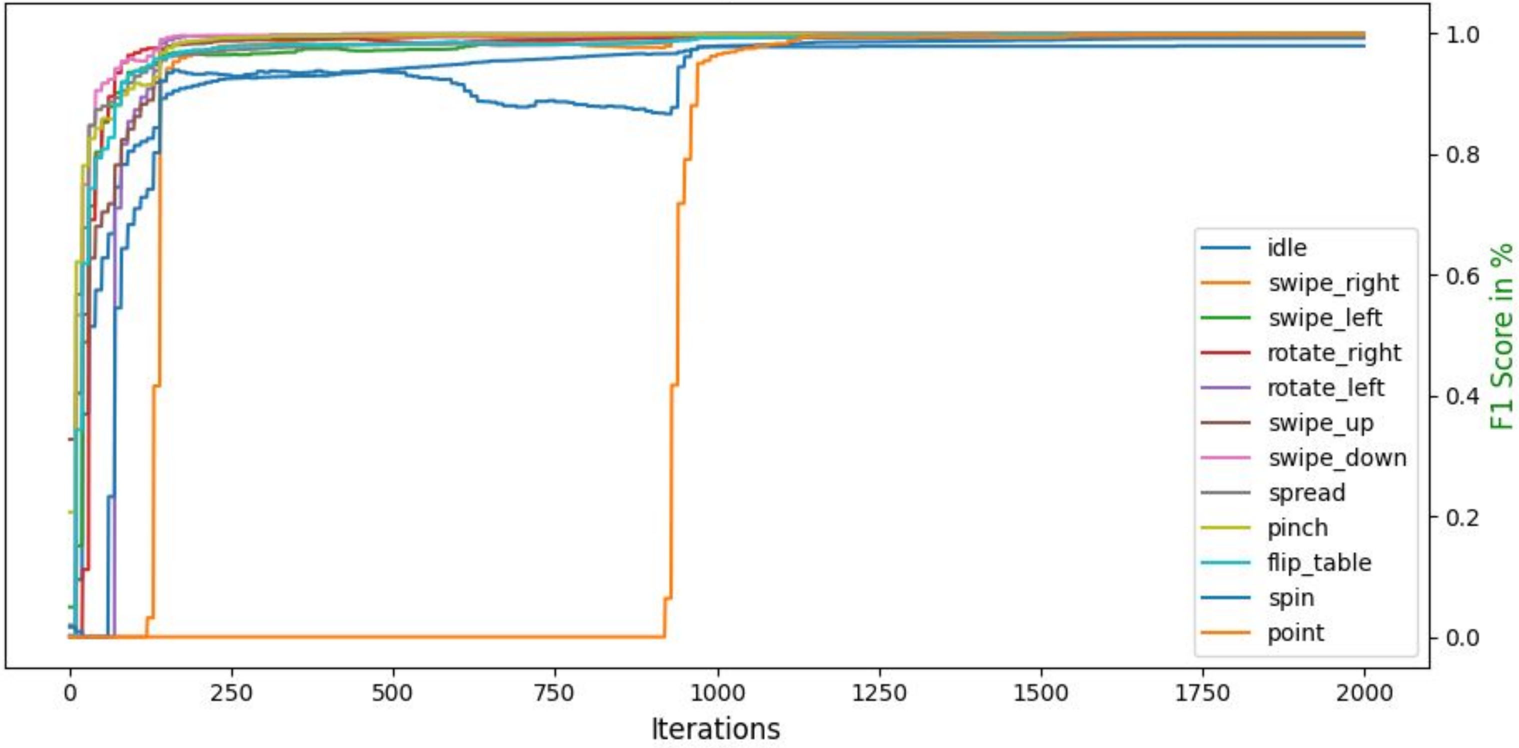


Validation: Precision and Recall per Class



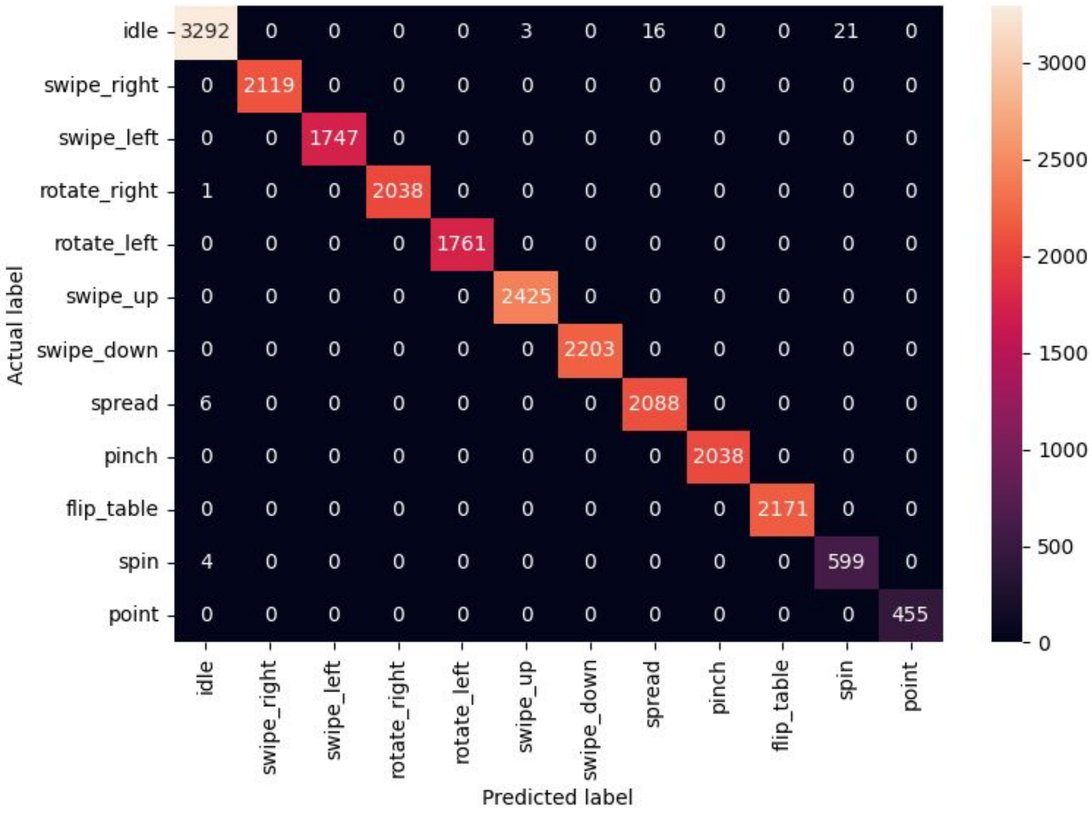


Validation: F1-Score per Class



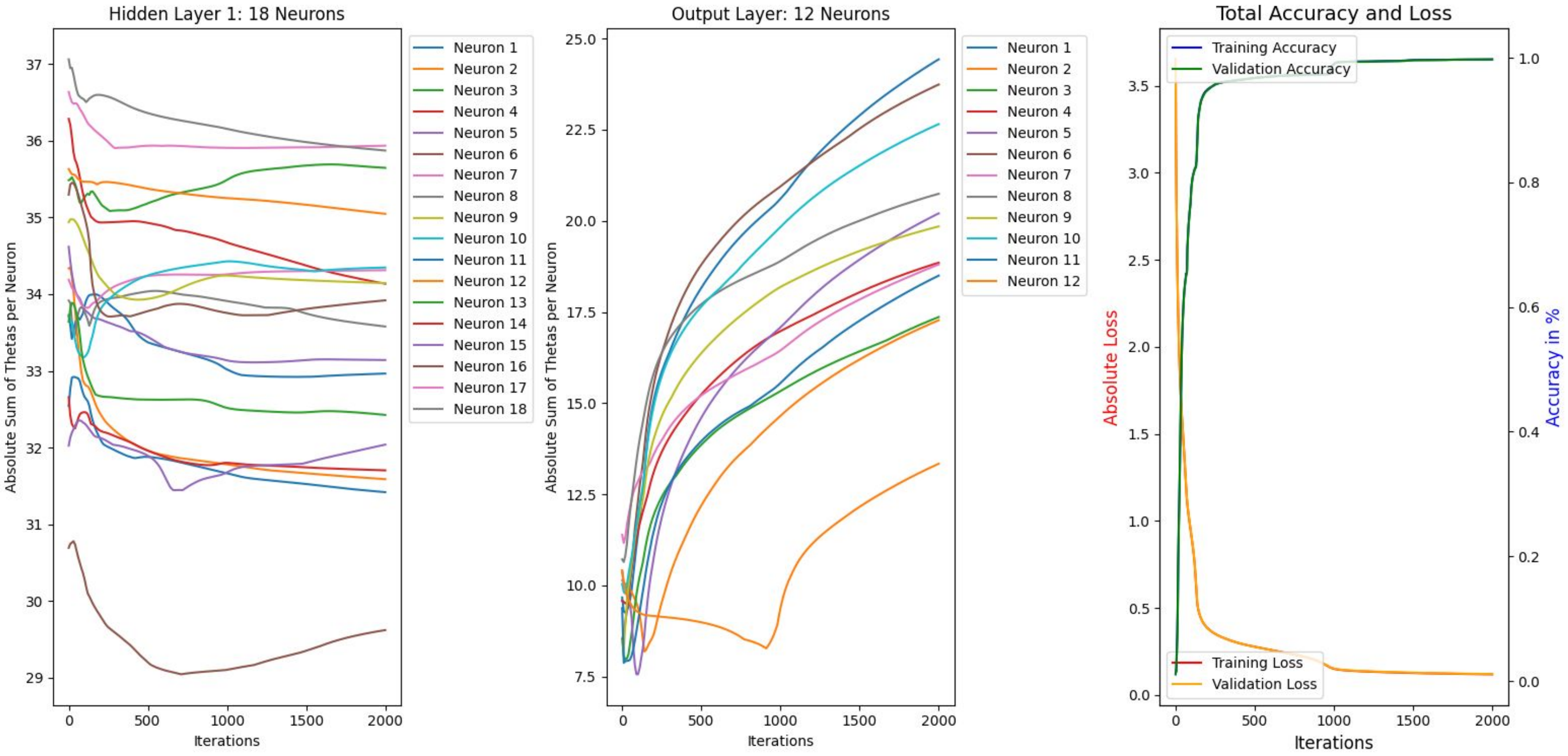
Validation: Confusion Matrix

Validation: Confusion Matrix



Actual label \ Predicted label	idle	swipe_right	swipe_left	rotate_right	rotate_left	swipe_up	swipe_down	spread	pinch	flip_table	spin	point
idle	3292	0	0	0	0	3	0	16	0	0	21	0
swipe_right	0	2119	0	0	0	0	0	0	0	0	0	0
swipe_left	0	0	1747	0	0	0	0	0	0	0	0	0
rotate_right	1	0	0	2038	0	0	0	0	0	0	0	0
rotate_left	0	0	0	0	1761	0	0	0	0	0	0	0
swipe_up	0	0	0	0	0	2425	0	0	0	0	0	0
swipe_down	0	0	0	0	0	0	2203	0	0	0	0	0
spread	6	0	0	0	0	0	0	2088	0	0	0	0
pinch	0	0	0	0	0	0	0	0	2038	0	0	0
flip_table	0	0	0	0	0	0	0	0	0	2171	0	0
spin	4	0	0	0	0	0	0	0	0	0	599	0
point	0	0	0	0	0	0	0	0	0	0	0	455

Training: Absolute Theta Sum per Neuron



Graphs are great, but let’s see it in action!