DATA SCIENCE PROJECT

- Zihao Zhang

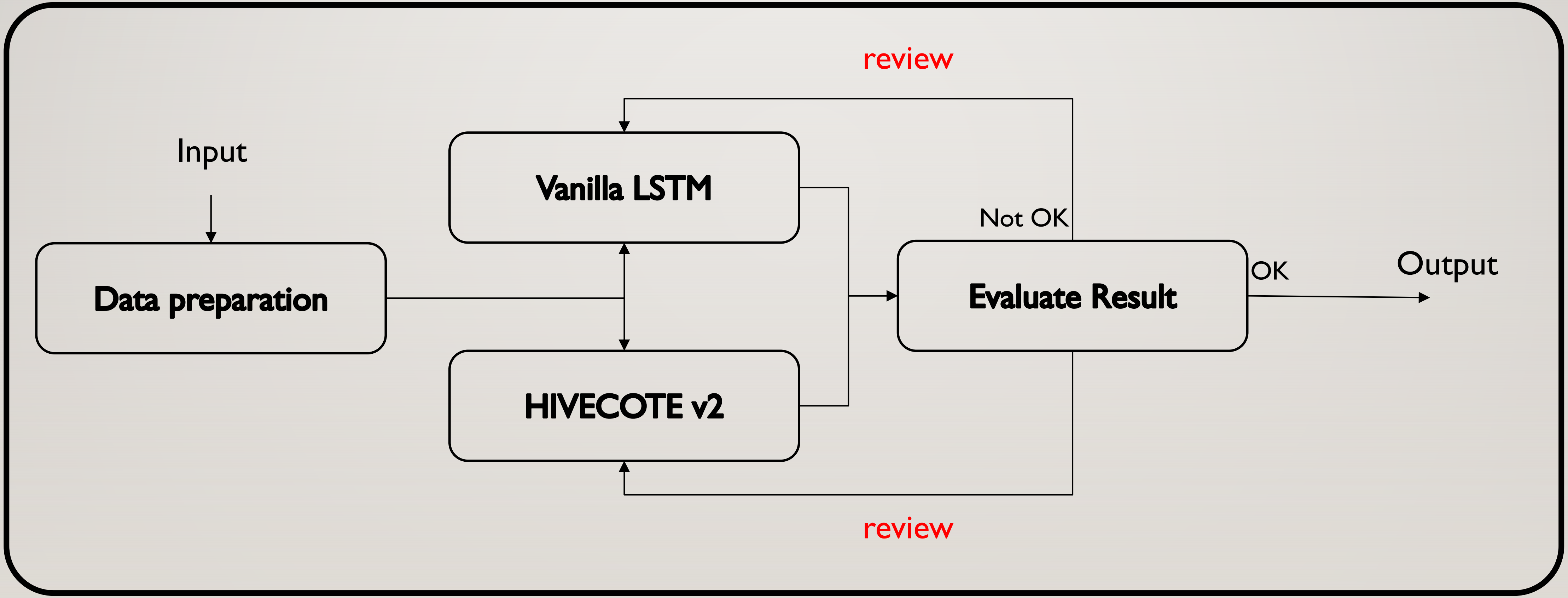
CONTENT

- Define the Problem
- Workflow
- Data Preparation
- Modelling
- Further Improvement

DEFINE THE PROBLEM

- What is the problem?
  - 370 samples belonging to 9 persons, classify 270 unknown samples
  - Each sample is a 7~29 * 12 dimensional time series (12 LPC)
  - Time series multi-class classification
- How to solve the problem?
  - Vanilla LSTM
  - SOTA model – HIVECOTE v2.0

WORKFLOW


review
Input
Vanilla LSTM
Not OK
Data preparation
Evaluate Result
OK
Output
HIVECOTE v2
review

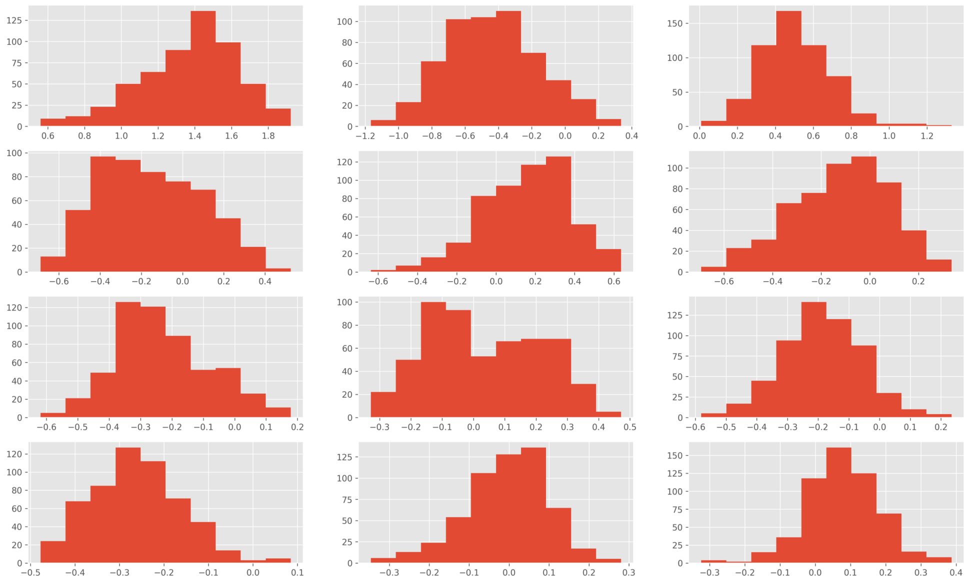
METRIC

- **Accuracy** for all classes
- Worst performance user's **accuracy**

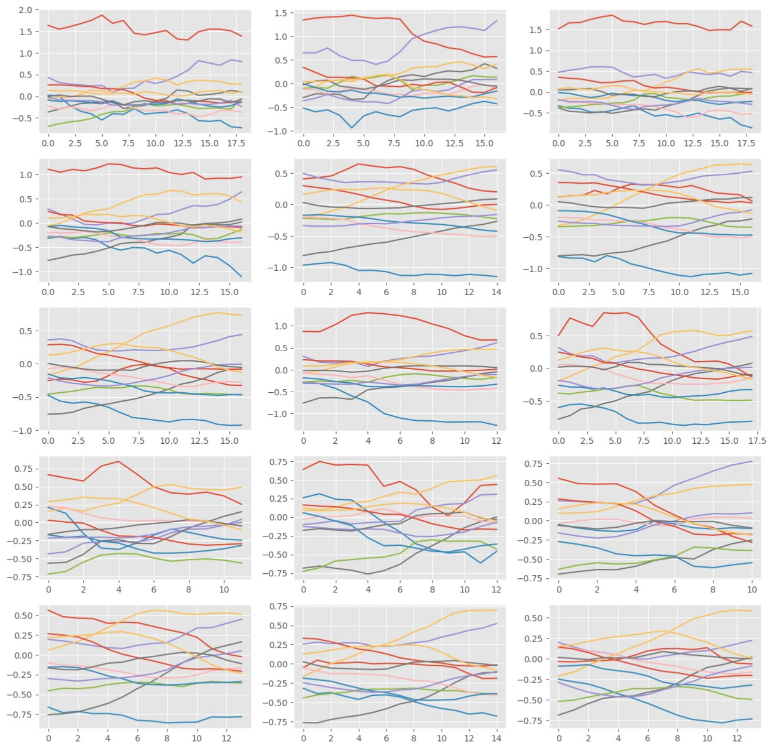
DATA PREPARATION

- Exploratory Data Analysis and Visualization

12 LPC histogram for person 1

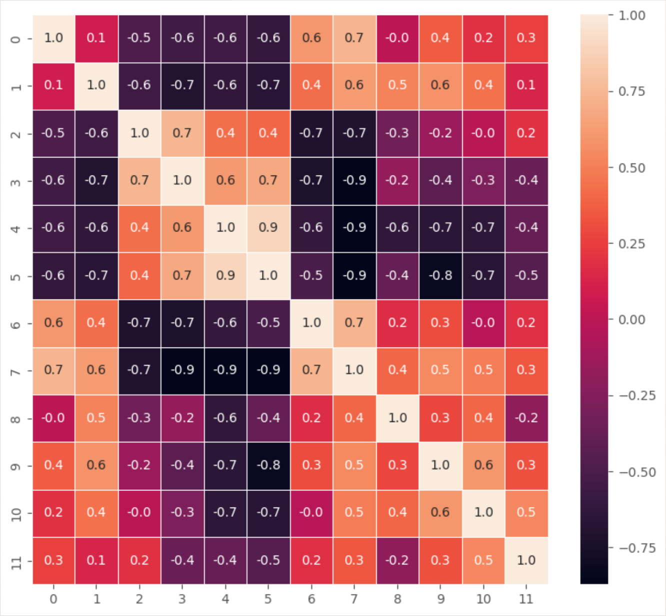


12 LPC line-plot for head 15 samples



DATA PREPARATION

- Exploratory Data Analysis and Visualization
  - High autocorrelation



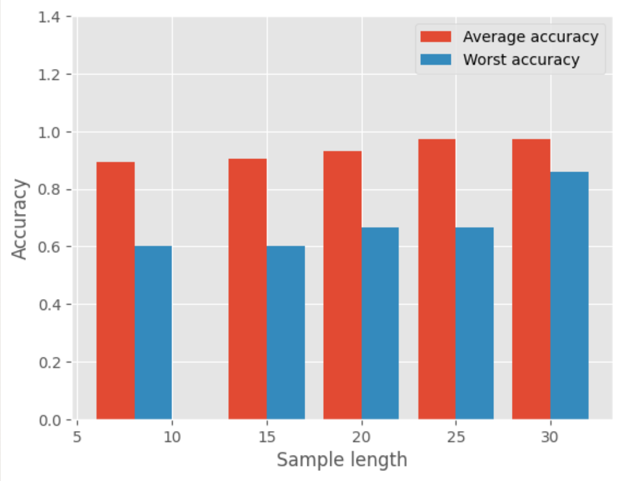
	0	1	2	3	4	5	6	7	8	9	10	11
0	1.0	0.1	-0.5	-0.6	-0.6	-0.6	0.6	0.7	-0.0	0.4	0.2	0.3
1	0.1	1.0	-0.6	-0.7	-0.6	-0.7	0.4	0.6	0.5	0.6	0.4	0.1
2	-0.5	-0.6	1.0	0.7	0.4	0.4	-0.7	-0.7	-0.3	-0.2	-0.0	0.2
3	-0.6	-0.7	0.7	1.0	0.6	0.7	-0.7	-0.9	-0.2	-0.4	-0.3	-0.4
4	-0.6	-0.6	0.4	0.6	1.0	0.9	-0.6	-0.9	-0.6	-0.7	-0.7	-0.4
5	-0.6	-0.7	0.4	0.7	0.9	1.0	-0.5	-0.9	-0.4	-0.8	-0.7	-0.5
6	0.6	0.4	-0.7	-0.7	-0.6	-0.5	1.0	0.7	0.2	0.3	-0.0	0.2
7	0.7	0.6	-0.7	-0.9	-0.9	-0.9	0.7	1.0	0.4	0.5	0.5	0.3
8	-0.0	0.5	-0.3	-0.2	-0.6	-0.4	0.2	0.4	1.0	0.3	0.4	-0.2
9	0.4	0.6	-0.2	-0.4	-0.7	-0.8	0.3	0.5	0.3	1.0	0.6	0.3
10	0.2	0.4	-0.0	-0.3	-0.7	-0.7	-0.0	0.5	0.4	0.6	1.0	0.5
11	0.3	0.1	0.2	-0.4	-0.4	-0.5	0.2	0.3	-0.2	0.3	0.5	1.0

DATA PREPARATION

- Padding: to make all data the same length
  - Maximum length, padding with last value – keep more information
  - Average length ×

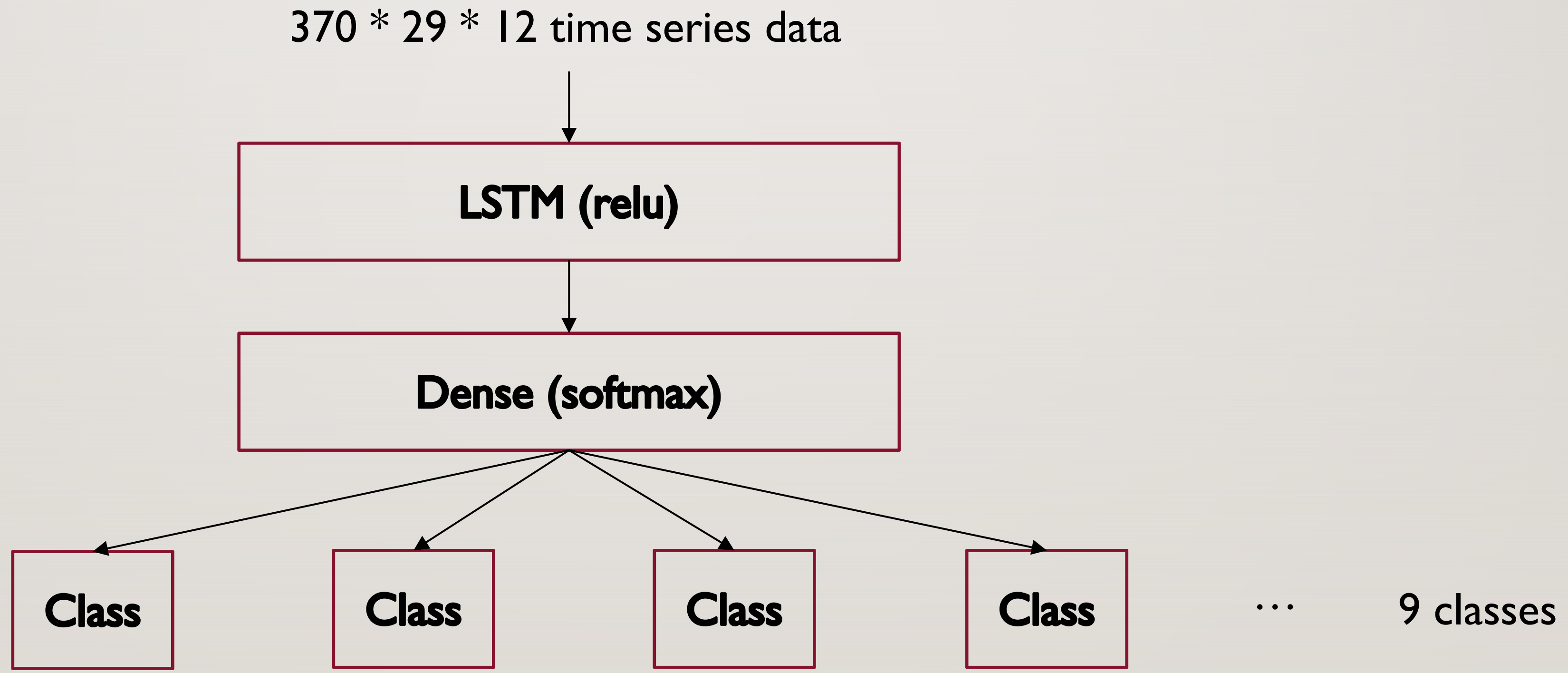
X_i	X_j	X_k
x_{i1}	x_{j1}	x_{k1}
x_{i2}	x_{j2}	x_{k2}
x_{i3}	x_{j3}	x_{k3}
x_{i4}	x_{j4}	x_{k4}
x_{i5}	x_{j5}	x_{k5}
x_{i5}	x_{j6}	x_{k6}
x_{i5}	x_{j7}	x_{k7}
x_{i5}	x_{j7}	x_{k8}

Maximum length



MODELLING

- Vanilla LSTM



370 * 29 * 12 time series data

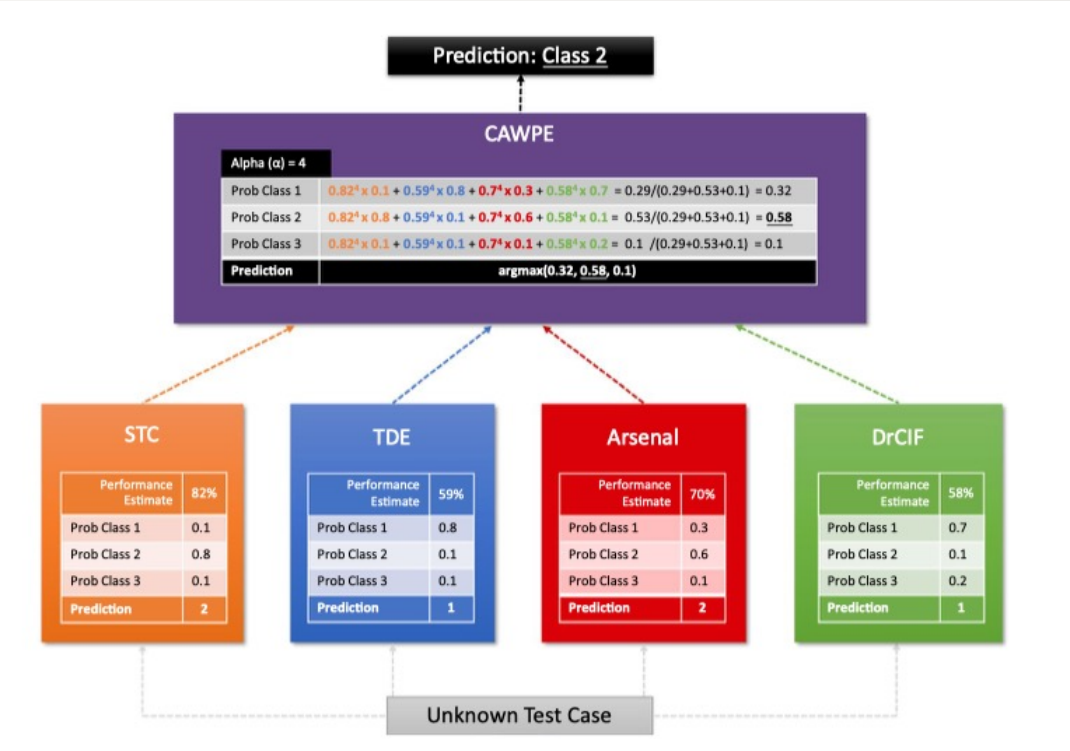
LSTM (relu)

Dense (softmax)

Class **Class** **Class** **Class** … 9 classes

HIVECOTE V2

- State of the art
- Ensemble model
  - Shapelet Transform Classifier
  - Arsenal
  - Dictionary based representation TDE
  - the interval based DrCIF



Prediction: Class 2

CAWPE

Alpha (α) = 4

Prob Class 1	$0.82^4 \times 0.1 + 0.59^4 \times 0.8 + 0.7^4 \times 0.3 + 0.58^4 \times 0.7 = 0.29/(0.29+0.53+0.1) = 0.32$
Prob Class 2	$0.82^4 \times 0.8 + 0.59^4 \times 0.1 + 0.7^4 \times 0.6 + 0.58^4 \times 0.1 = 0.53/(0.29+0.53+0.1) = \underline{0.58}$
Prob Class 3	$0.82^4 \times 0.1 + 0.59^4 \times 0.1 + 0.7^4 \times 0.1 + 0.58^4 \times 0.2 = 0.1\ /(0.29+0.53+0.1) = 0.1$
Prediction	argmax(0.32, <u>0.58</u>, 0.1)

STC

Performance Estimate	82%
Prob Class 1	0.1
Prob Class 2	0.8
Prob Class 3	0.1
Prediction	2

TDE

Performance Estimate	59%
Prob Class 1	0.8
Prob Class 2	0.1
Prob Class 3	0.1
Prediction	1

Arsenal

Performance Estimate	70%
Prob Class 1	0.3
Prob Class 2	0.6
Prob Class 3	0.1
Prediction	2

DrCIF

Performance Estimate	58%
Prob Class 1	0.7
Prob Class 2	0.1
Prob Class 3	0.2
Prediction	1

Unknown Test Case

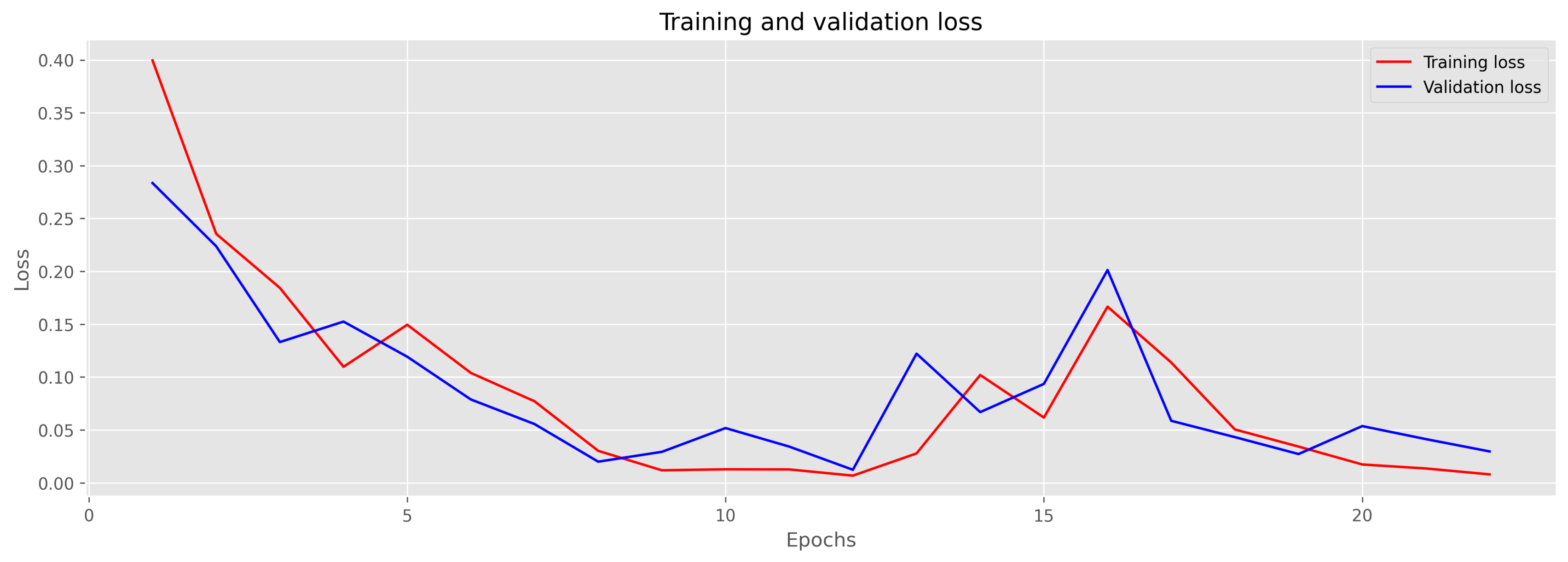
Picture source: https://arxiv.org/pdf/2104.07551.pdf

IMPROVE THE MODEL

- Augmentation:
  - Weighted resample augmentation
  - All dataset augmentation
  - One class resample augmentation

Yield best results

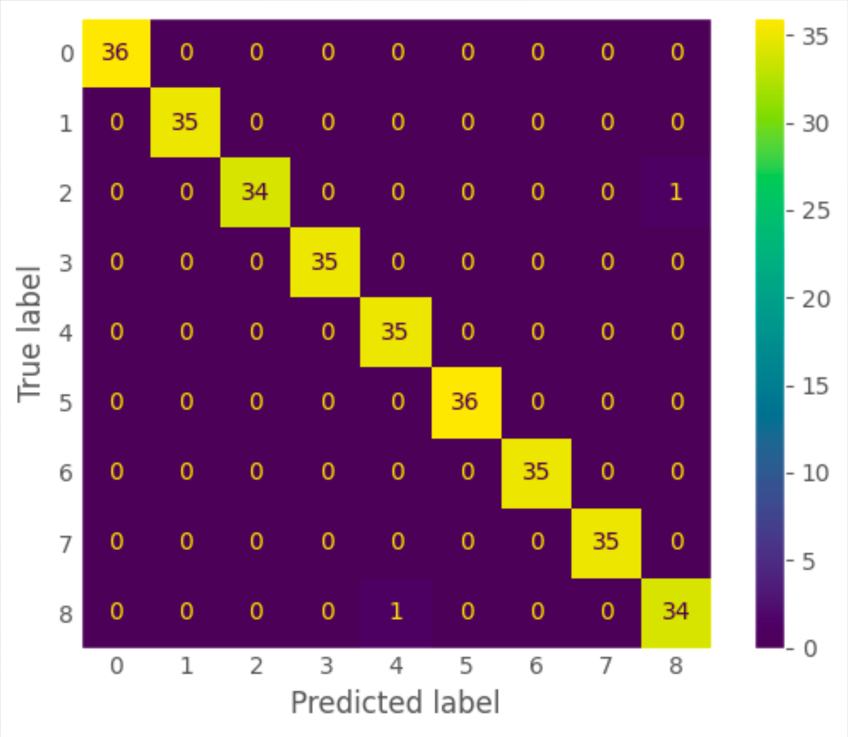
VANILLA LSTM



BEST validation set performance:

- The classification accuracy is 0.9874
- The worst performing user accuracy is 0.9429

HIVECOTE V2



	0	1	2	3	4	5	6	7	8
0	36	0	0	0	0	0	0	0	0
1	0	35	0	0	0	0	0	0	0
2	0	0	34	0	0	0	0	0	1
3	0	0	0	35	0	0	0	0	0
4	0	0	0	0	35	0	0	0	0
5	0	0	0	0	0	36	0	0	0
6	0	0	0	0	0	0	35	0	0
7	0	0	0	0	0	0	0	35	0
8	0	0	0	0	1	0	0	0	34

Confusion matrix

BEST validation set performance:

- The classification accuracy is 0.9937
- The worst performing user accuracy is 0.9714

RESULT IMPROVEMENT

- Final result is not robust and worst user prediction accuracy vibrates (best save to csv) add regularization
- More feature engineering
- Gather more data from the last person
- Better data augmentation methods (slicing, warping, jittering, rotation, and their combination)
- Hyperparameters tuning
- HIVECOTE v2 is slow, especially for large dataset (replaced by RocketClassifier)
- Data Leakage problem
- Focal Loss (add BinaryFocalCrossentropy as the evaluation criteria)

- DO NOT DO
  - Include test data with labels (easy to guess) for training ☺

THANK YOU!

- Zihao Zhang