# Top-1 CORSMAL challenge 2020 submission

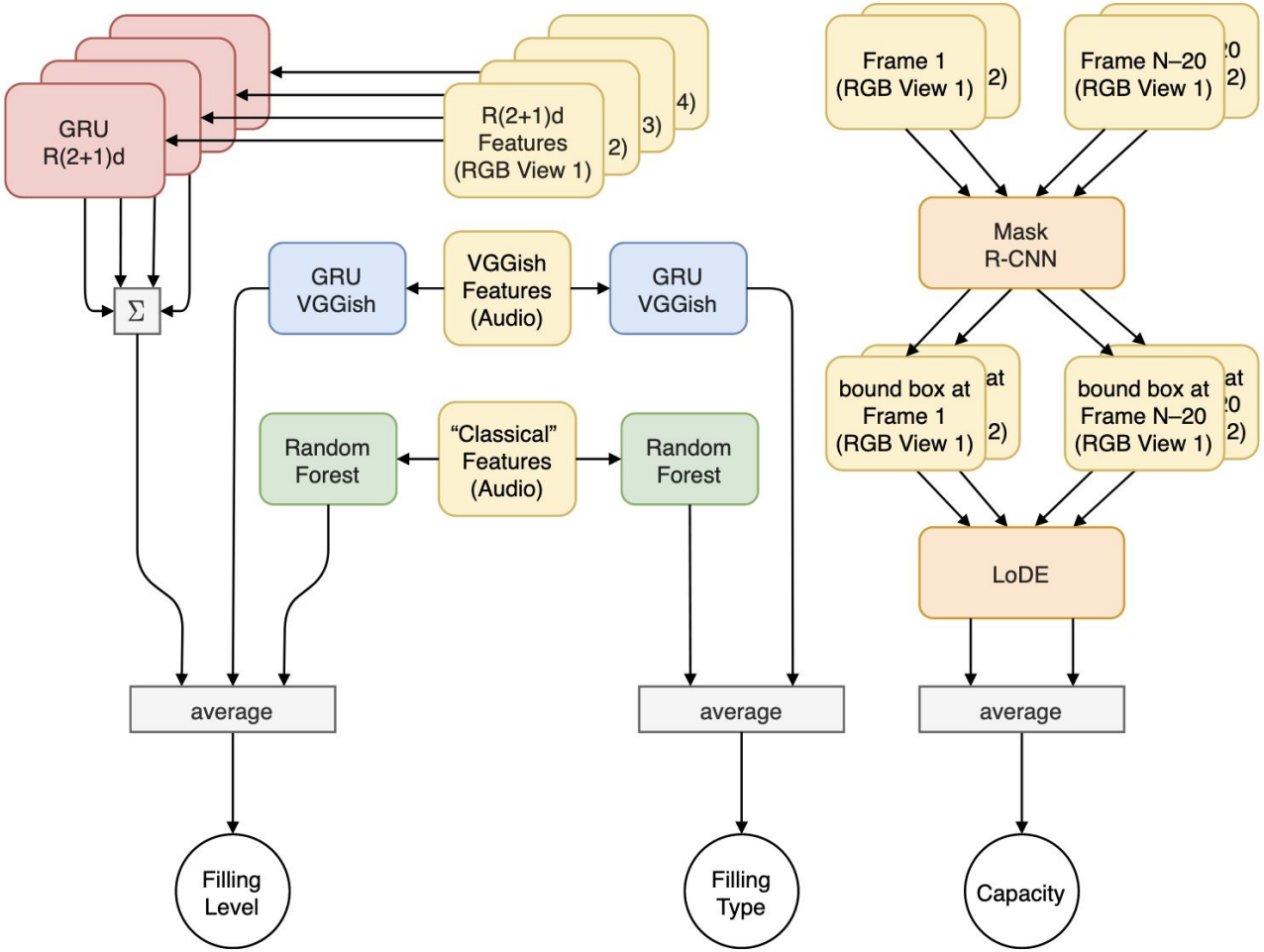
Filling mass estimation using multi-modal observations of human-robot handovers

**Vladimir Iashin** Francesca Palermo
Gökhan Solak Claudio Coppola

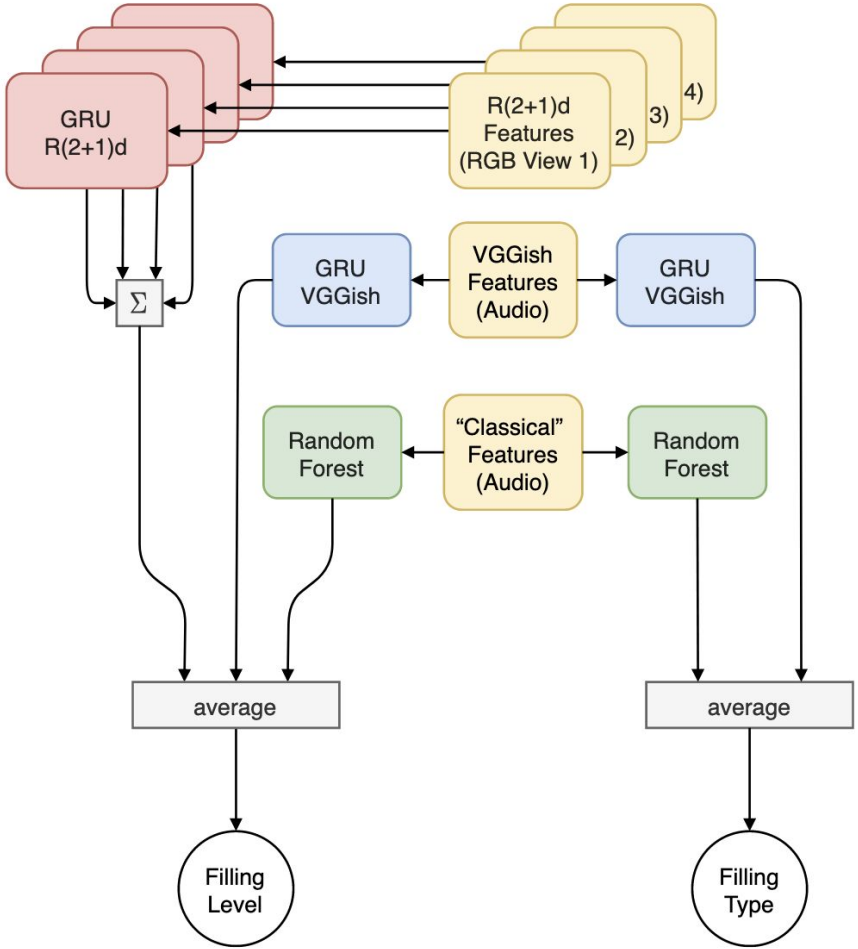
# Motivation

- Robotic manipulation is widely used in heavy industry
- Yet it remains to be a research area for domestic robotics
- Handover is one of the challenges for domestic robotics
- Filling mass estimation is a key challenge for handover

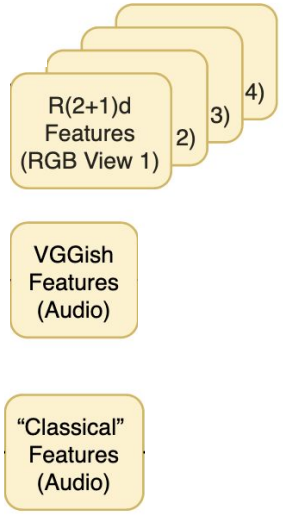
# Overview


GRU
R(2+1)d
R(2+1)d
Features
(RGB View 1)
Σ
GRU
VGGish
VGGish
Features
(Audio)
GRU
VGGish
Random
Forest
"Classical"
Features
(Audio)
Random
Forest
average
average
Filling
Level
Filling
Type
Frame 1
(RGB View 1)
Frame N–20
(RGB View 1)
Mask
R-CNN
bound box at
Frame 1
(RGB View 1)
bound box at
Frame N–20
(RGB View 1)
LoDE
average
Capacity

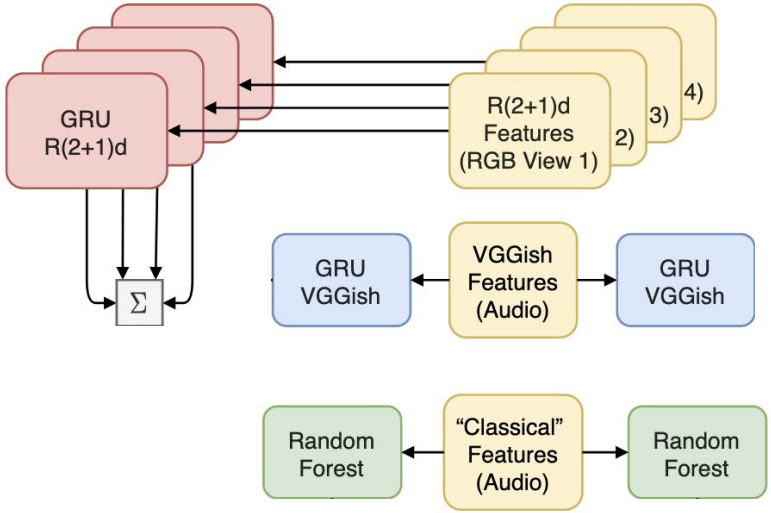
# Filling Level & Filling Type Classification



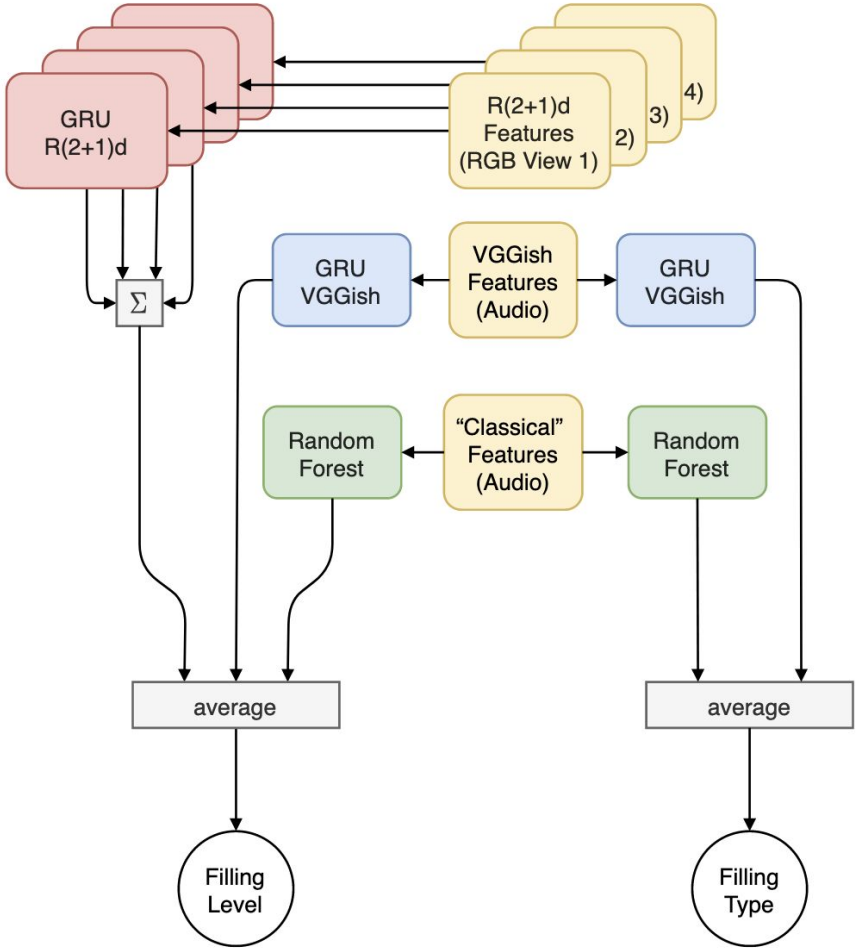
# Feature Extraction



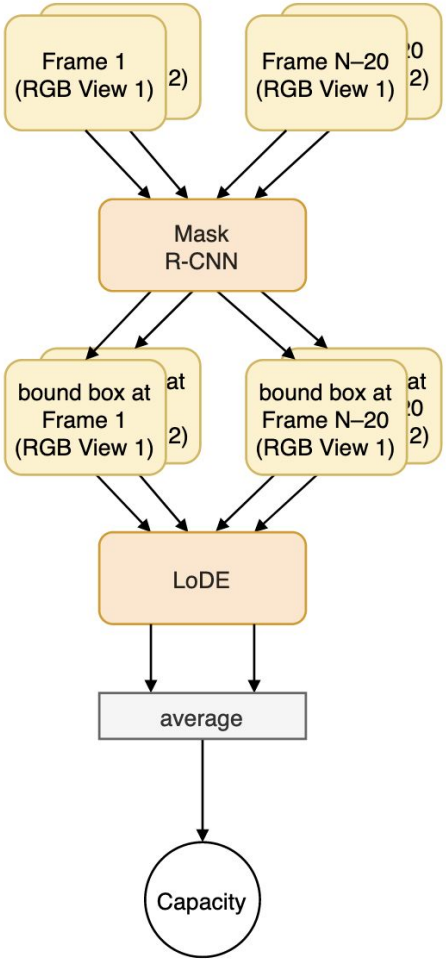
# Classification Models



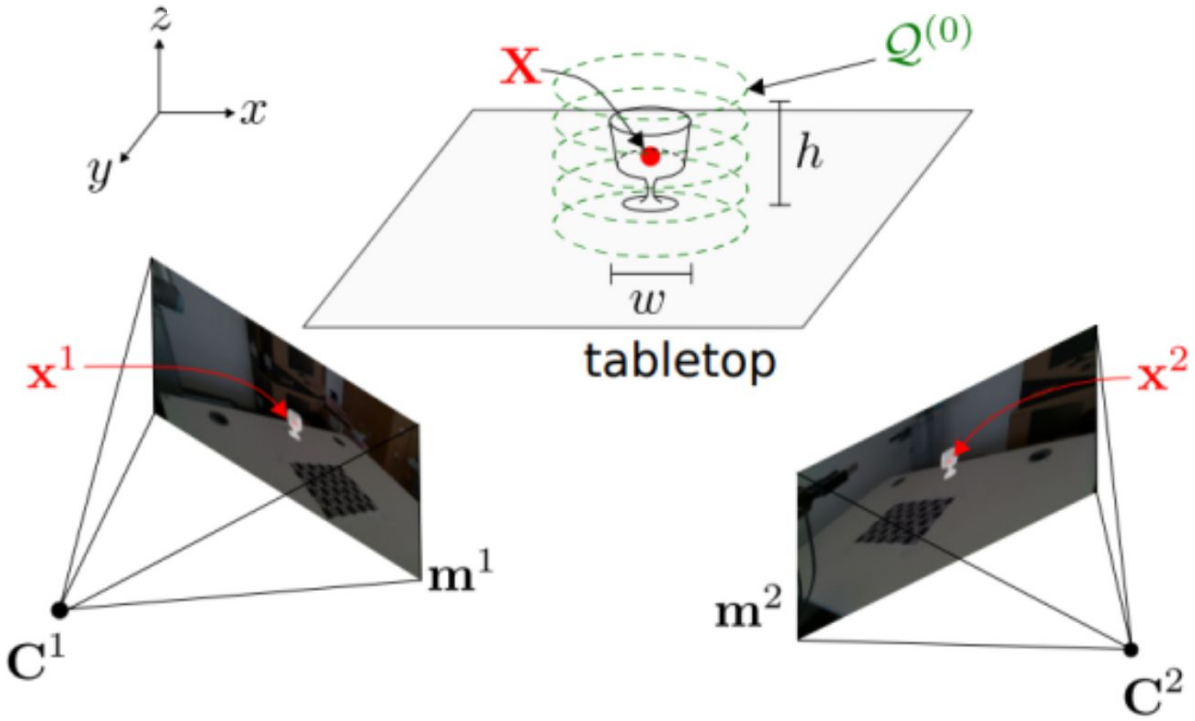
# Result Aggregation



# Capacity Estimation



$$C = \overline{r}^2 \cdot h \cdot \pi$$



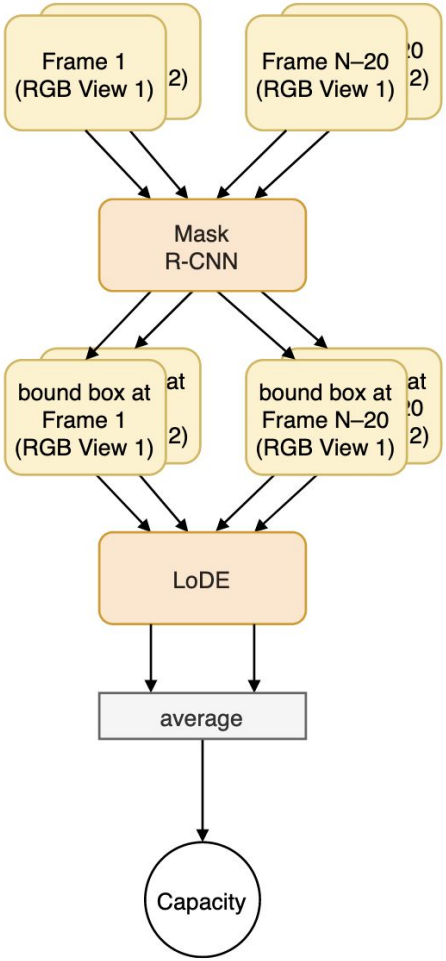
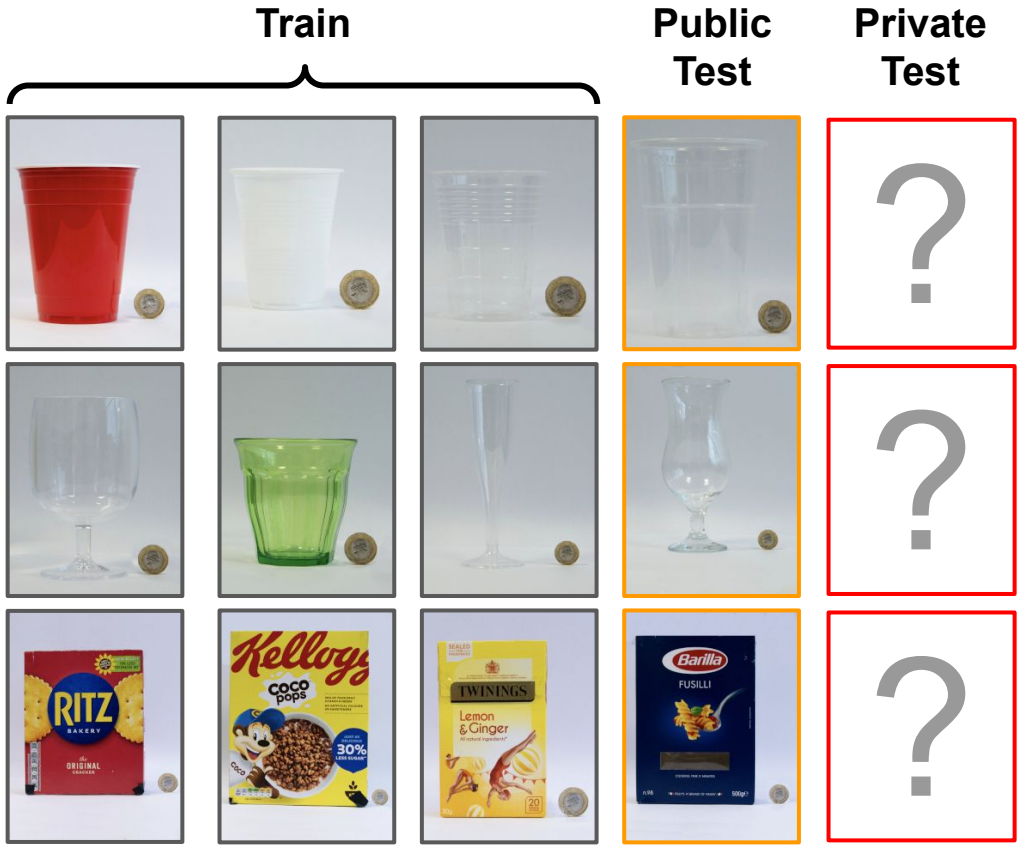




Image was borrowed from the original publication

# Dataset (CORSMAL 2020)

- 15 containers:
  - 5 drinking **cups**,
  - 5 **glasses**,
  - 5 **food boxes**
- Filling Level Classes
  - 0 %
  - 50 %
  - 90 %
- Filling Type Classes
  - Rice
  - Pasta
  - Water (for glasses and cups)
  - Empty
- In total, 1140 events



# Implementation Details

github.com/v-iashin/CORSMAL

github.com/v-iashin/CORSMAL

# Cross-validation Approach

Fold #1


Train Train Valid

Fold #2


Train Valid Train

Fold #3


Valid Train Train

# Results

| | Validation Set | | |
|---|---|---|---|
| **Sub-task** | **"class." feats.** | **VGGish** | **R(2+1)d** |
| Filling Level | 69.9 | 75.5 | 74.7 |
| Filling Type | 93.3 | 91.3 | –* |

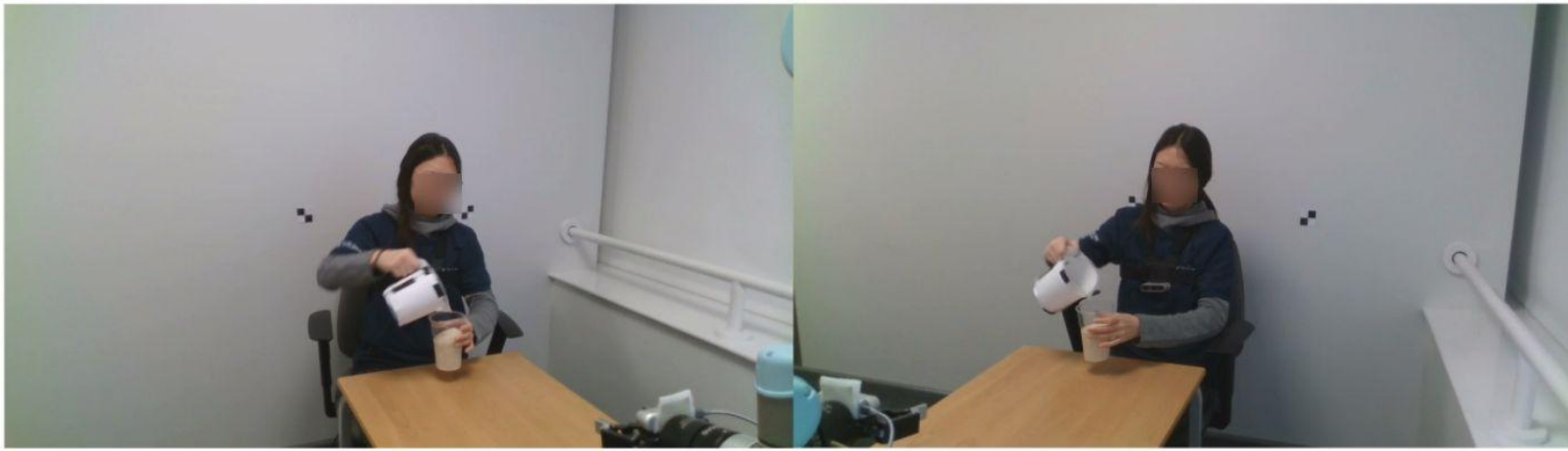
* Yes, we tried. F1 = 67.3

Metric: Weighted F1 averaged across 3 validation folds

C1 Camera View (Left)

C2 Camera View (Right)

Results (Capacity Estimation)



# Result on Test Sets

| Team | Description | Task 1 | Task 2 | Task 3 | Public | Private | Overall |
|---|---|---|---|---|---|---|---|
| Because It's Tactile | GRU+ Random Forest for filling properties estimation. LoDE with RGB-D-IR data from selected frames in a video for volume estimation. | ✓ | ✓ | ✓ | 64.98 | 65.15 | 65.06 |
| HVRL | Log-Mel spectrogram-based audio features as input to VGG-based CNN and LSTM for filling properties estimation. Container volume from the shape approximation as cuboid of the 3D point cloud obtained with RGB-D data and object detection with Mask R-CNN. | ✓ | ✓ | ✓ | 63.32 | 61.01 | 62.16 |
| Concatenation | Multi-modal learning with audio features and prior of container categories through object detection for inferring container capacity and fluid properties. | ✓ | ✓ | ✓ | 52.80 | 54.14 | 53.47 |
| NTNU-ERC | MFCC features in a 20s-window + neural network to classify filling type. Object detection and selection of the closest contours (up to 700 mm) in the depth data + regression with a CNN for container capacity. | | ✓ | ✓ | 38.56 | 39.80 | 39.18 |
| Random | Baseline with random estimations for each task. | ✓ | ✓ | ✓ | 38.47 | 31.65 | 35.06 |
| Challengers | Sound-based classification of filling type and level with STFT and 5-layers fully connected neural network. | ✓ | ✓ | | 29.25 | 23.21 | 26.23 |
| SCC-Net | Sound-based hierarchical ensemble of DNNs to jointly classify filling type and level. | ✓ | ✓ | | 28.02 | 22.92 | 25.47 |

| | Test set | |
|---|---|---|
| **Sub-task** | **Public** | **Private** |
| Filling Level | 78.14 | 81.16 |
| Filling Type | 93.83 | 94.70 |
| Container Capacity | 60.56 | 60.58 |
| Overall Performance | 64.98 | 65.15 |

# Source Code

Claudio Coppola

Gökhan Solak

Francesca Palermo

Vladimir Iashin