CNN for Hand Washing Movement Classification: What Matters More – the Approach or the Dataset?

Atis Elsts[1,2], Maksims Ivanovs[1], Roberts Kadikis[1] and Olegs Sabelnikovs[2]

[1] Institute of Electronics and Computer Science (EDI), Riga LV-1006, Latvia
[2] Medical Education Technology Centre (METC), Riga Stradins University, Riga LV-1007, Latvia

***Abstract*—Good hand hygiene is one of the key factors in preventing infectious diseases, including COVID-19. Advances in machine learning have enabled automated hand hygiene evaluation, with research papers reporting highly accurate hand washing movement classification from video data. However, existing studies typically use datasets collected in lab conditions. In this paper, we apply state-of-the-art techniques such as MobileNetV2 based CNN, including two-stream and recurrent CNN, to three different datasets: a good-quality and uniform lab-based dataset, a more diverse lab-based dataset, and a large-scale real-life dataset collected in a hospital. The results show that while many of the approaches show good accuracy on the first dataset, the accuracy drops significantly on the more complex datasets. Moreover, all approaches fail to generalize on the third dataset, and only show slightly-better-than random accuracy on videos held out from the training set. This suggests that despite the high accuracy routinely reported in the research literature, the transition to real-world applications for hand washing quality monitoring is not going to be straightforward.**

***Index Terms*—CNN, movement classification, hand washing, hand hygiene**

I. Introduction

Good hand hygiene is one the most important factors in preventing transmission of germs and associated infections, including COVID-19. The World Health Organization (WHO) has published recommended practices [10] for hand washing and alcohol-based hand rubbing, both of which describe the same six main hand washing steps required to ensure thorough cleaning of hands. Unfortunately, even medical professionals often fail to observe these guidelines, leading to a large number of hospital-transmitted infections.

Automated hand-washing quality monitoring is therefore urgently needed to improve compliance and to prevent these infections. A key task here is to be able to recognize all six key washing movements in order to detect whether they all have been executed. This has been an active area in the recent years, with multiple papers reporting high hand washing movement recognition accuracy using CNN classifiers [2], [7], [8]. These papers typically use datasets collected in a lab, such as the Kaggle challenge dataset [1], and utilize CNN based on pre-trained models such as MobileNetV2, extending it with a multi-stream network architecture, or incorporating recurrent elements such as LSTM.

This research is funded by the Latvian Council of Science project: "Automated hand washing quality control and quality evaluation system with real-time feedback", No: lzp-2020/2-0309.

Our goal is to apply these state-of-the-art techniques to more complex datasets, with emphasis on-light weight classifiers that can be run on mid-range smartphones rather than require powerful hardware accelerators. To achieve this goal, we formulate five intuitive hypotheses:

- *H1:* It is possible to successfully apply classifiers suitable for the lab-collected Kaggle dataset to other datasets, specifically: (1) to a more diverse lab-collected dataset, and (2) to a real-life dataset.
- *H2:* Adding temporal information through optical flow or time-distributed inputs increases classification accuracy.
- *H3:* Using extra layers increases classification accuracy, at the cost of additional training and inference duration.
- *H4:* A full retraining of the base model increases classification accuracy, at the cost of additional training duration.
- *H5:* Out-of-the-box Keras data augmentation layers are sufficient for good generalization performance.

We investigate the hypotheses using three different CNN architectures[1]:

1) Baseline MobileNetV2 CNN with RGB inputs;
2) Two-stream MobileNetV2 based CNN with RGB and optical flow inputs;
3) Recurrent MobileNetV2 based CNN with GRU elements and time-distributed RGB inputs.

In order to do that, we train and evaluate the CNN on three different datasets:

1) The Kaggle challenge dataset [1];
2) A lab-collected dataset from METC [5];
3) A large-scale real-life dataset from Pauls Stradins Clinical University Hospital [6].

Unfortunately, the results (Section IV) show evidence against all five of these hypotheses. For example, techniques that show good performance on the "simple" lab dataset show mediocre results on the more complex one, and fail to generalize on the real-life dataset. This outcome suggests that it is not straightforward to scale the approaches from the current research literature to real-life applications, and that the published results in this area should be taken with a grain of salt. This negative result is the main scientific contribution of this paper.

The paper is structured as follows: we overview related work in Section II; describe our approach in Section III, present results in Section IV, and summarize the results in Section V.

[1]The code used in our experiments is available at https://github.com/edi-riga/handwash.

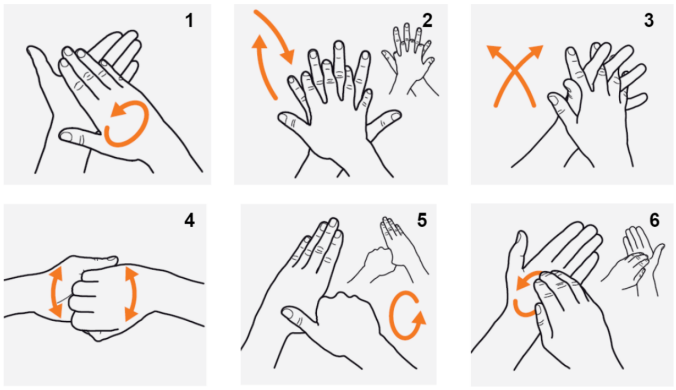


Fig. 1: The six key movement of hand washing [10].

II. Related Work

Deep neural networks have demonstrated state-of-the-art results in many image classification tasks [4], including applications in hand hygiene monitoring [11], [12]. There has been a surge of interest in attempts to classify the hand washing movements according to the 6-step guide by the WHO (Fig. 1).

Prakasa & Sugiarto [8] extract frames from videos, convert the resulting images from red, green, and blue (RGB) channels to hue, saturation, and value (HSV) channels to obtain image component with high contrast on the skin human region (the hue channel), and classify them using a custom CNN classifier. However, the dataset they use consists of a single instructional video, therefore the generalization accuracy is not adequately tested.

Nagaraj *et al.* [7] design a three-stream network architecture, based on a classical papers on two-stream CNN fusion [3], [9]. The three streams utilize RGB frames, optical flow frames, and histogram of gradients as the inputs, in this way incorporating spatial, temporal, as well as object-level information from the videos. The authors use the full Kaggle dataset [1] to show that their approach performs better of than any of the three modalities used alone, and achieve 86.6 % accuracy on the hand wash dataset. We note that accuracy is measured for 12-class separation, and is likely to be above 95 % if just 7 movement classes were considered, as we do in the present paper. (The 12-class problem arises when left-hand and right-hand washing movements are treated as separate classes.) The authors provide a GitHub repository with an implementation of the fusion classifier.

Another recent study by Cikel *et al.* [2] uses the publicly available subset of the Kaggle dataset [1] with hand-washing movements to train and evaluate 3 models consisting of a Resnet-152 CNN encoder and a decoder based on a 3-layer LSTM, using as input the RGB frames of the videos for the first one, the optical flow for the second one, and a two-stream input made up of both RGB frames and optical flow for the third one. The RGB network achieves an accuracy of 97.33%.

Our own work on the hospital dataset [6] introduces the dataset itself, as well as reports 75 % classification accuracy from some initial experiments. However, the F_1 score of those results is lower than 0.75 (though still above 0.5) because of some class imbalance. Most importantly, this previous work does not attempt to measure the generalization performance of the classifier by evaluating it on users and washing locations that are not part of the training data.

III. Methods

A. Datasets

We use three datasets in our work (Table I). One representative image from each datasets is shown in Fig. 2.

The "Kaggle" dataset used in this work is downloaded from the Kaggle "Hand Wash Dataset" [1] challenge page. The complete dataset has 292 hand washing episodes; however, we use the freely available part of it, with just 25 episodes. Each episode has high-quality scripted hand washing videos corresponding to each of the hand washing steps defined by the WHO [10].

The METC dataset was collected in July–August 2021, as part of a user feedback evaluation study [5]. The main goal of the experiments is to investigate the effect from mobile application based feedback during the washing process. The videos were recorded in multiple sessions with multiple users, but all experiments took place in the same location (i.e. had the same sink). Most of the users were medical students with previous knowledge about the expected hand washing steps, and were shown a reminder before the study, as well as visual aids on the smartphone screen during the washing procedure. However, while were knowledgeable and were instructed to complete the task to the best of their ability, imperfect execution was still present. A few users did not even complete all six washing movements.

The hospital dataset shows medical staff washing their hands as part of their normal job duties. These videos from real-life conditions include hand washing positions that are partially out out the frame or partially occluded, as well as low and variable lightning conditions. In this data, Imperfect and incomplete execution of the washing steps is a rule rather than an exception.

The data in the METC and hospital datasets is annotated according to the hand hygiene guidelines from the WHO [10], which identifies the six key movements (Fig.1). The movements and hand positions that do not corresponds to any of these six are labeled with the code 0 ("other").

TABLE I: Datasets.

Parameter / DS	**Kaggle**	**METC**	**Hospital**
Washing episodes	25	213	3185
Users	$\leq$25	71	many
Locations	$\leq$25	1	9
Environment	Lab	Lab	Real-life
Frame dimensions	720x480	640x480	640x480, 320x240
FPS	30	16	30

(a) Kaggle dataset

(b) METC dataset

(c) Hospital dataset

Fig. 2: Example images from the three datasets, with movement class label "1".

B. Data preparation

Some classes in the **Kaggle dataset** is merged so that left-hand and right-hand movements belong to the same class. The wrist washing movement ("step 7") is not labeled of the other datasets, so it is treated as the "other" movement (class 0).

The data from the **METC dataset** is separated in segments with continuous sequence of frames that all have the same class label. As the dataset is annotated in real-time, by a human operator during the data collection experiments [5], there is some reaction time that needs to be discounted. For this purpose, we remove a 1 second long video segment every time the class label changes in the data-stream.

The **hospital dataset** data is similarly separated in continuous sequences. This dataset has multiple annotators for most videos. Only parts of the dataset where two or more annotators have assigned matching class labels are used for CNN training and evaluation. We do not attempt to preprocess the data to improve the image quality, normalize the light levels, etc.

While the initial resolution of the videos is different, before using it as inputs to the CNN models, all frames are scaled down to 320x240 pixels using the standard Tensorflow functionality (with "interpolation" as the resize method).

Each dataset is split in two parts; one part (30 %) is used as the test data, and the other part as training and validation data. The METC and the hospital datasets have new, previously unseen users included as part of the test data. The hospital dataset comes with location information for each video. We split hospital dataset so that the test subset has videos from different locations than the training & validation subset.

Random sampling and class weighting is performed to account for the disbalance in some of the datasets. First, in the hospital dataset, a random portion of the data with class label 0 is dropped to increase the training speed of the classifiers. Subsequently, class weighting is used to deal with the remaining imbalance in all datasets.

Finally, the data is augmented with random flips and rotations during the training stage in order to improve the generalization performance of the classifiers.

C. Architectures

We use the Keras framework and select the pretrained MobileNetV2 model for baseline performance measurements

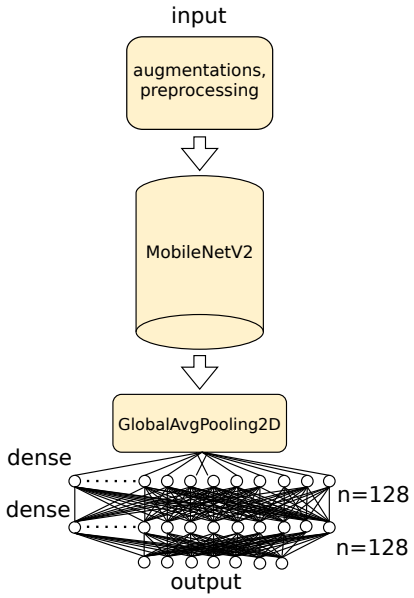


Fig. 3: Baseline CNN architecture.

(Fig. 3). MobileNetV2 architecture is selected because it is lightweight and has a good performance / accuracy tradeoff in many applications. The weights of the MobileNetV2 model are pretrained on the ImageNet dataset. The classifier takes RGB images as inputs; see Table II for details. On top of the base line classifier, one to three dense layers are deployed. The top layer has 7 units and uses the `softmax` activation function. In addition to this baseline, we investigate these more complex architectures:

- Two-stream network (Fig. 4), with two MobileNetV2 models in the base, joined by a fusion layer, and with one or more dense layers on top of the fusion layer;
- Recurrent CNN (Fig. 5), with a time-distributed layer uniting a number of base models, with GRU used as the memory unit.

The baseline network only takes a single frame as the input, while the more complex architecture also utilize temporal information as part of their inputs. The existing literature [7] argues that temporal information is required for accurate

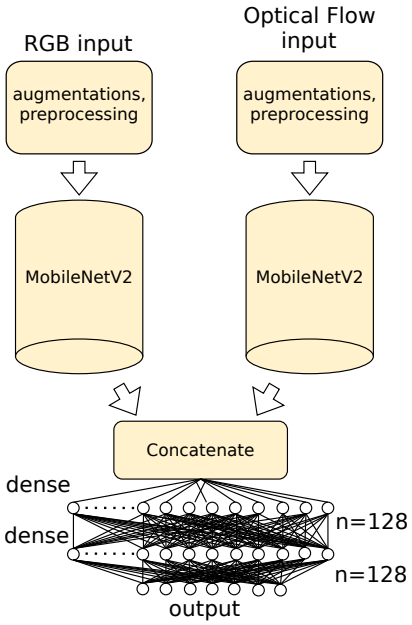


Fig. 4: Two-stream CNN architecture.

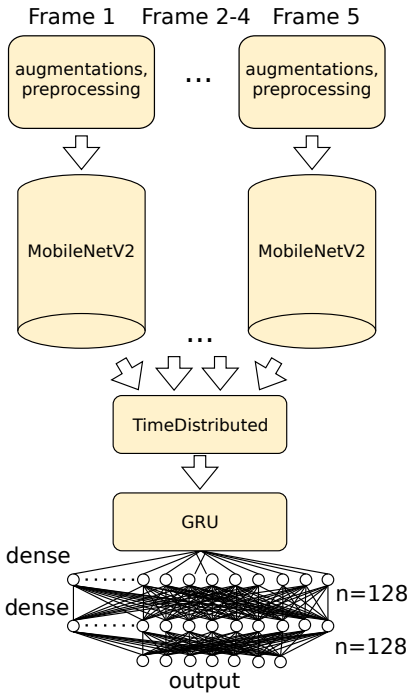


Fig. 5: Recurrent CNN architecture.

TABLE II: Neural network parameters.

Parameter	Value
All networks / default values	
Base model	MobileNetV2
Initial weights	ImageNet, 224x224
Input image dimensions	320x240x3
Data augmentations	Rotations, flips
Num. dense layers	1
Layers retrained	1 (“top”) or all (“full”)
Num. epochs	20
Batch size	32
Optimizer	Adam
Loss function	Cross entropy
Num. classes	7
Two-stream networks	
Streams	RGB & optical flow
Fusion	Before dense layers
Optical flow type	Farneback
Optical flow step	0.33 sec
Recurrent networks	
Recurrent element	GRU
Frame step	0.2 sec
Num. frames	5
Extra layer networks	
Num. dense layers	3
Transfer learning networks	
Num. epochs	10

movement recognition, as it is not possible to differentiate between the movement 1 and movement 3 from a single image.

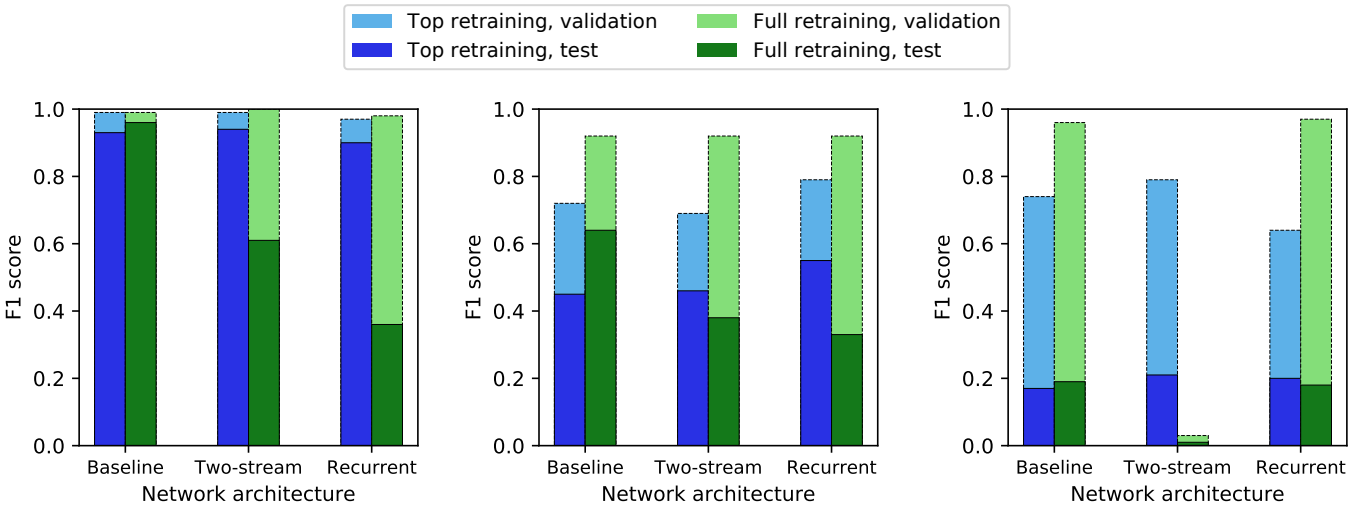
D. Additional Experiments

We perform two following additional experiments using the architectures as above.

- Extend the top of the network with two additional dense layers with 128 neurons each. A dropout of 0.2 is used after the two intermediate dense layers.
- Transfer learning: measure the generalization performance across datasets, before and after 10 epochs of fine tuning. Only generalizations from the less complex to more complex datasets are investigated.

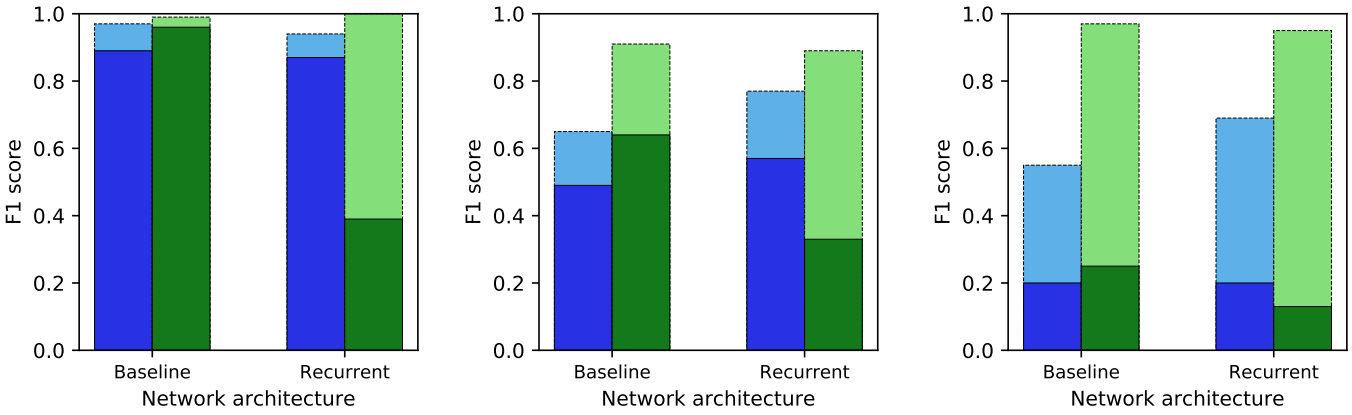
Finally, we also peform initial work on investigated a few more approaches. As these additional experiments fail to show notable improvements on the results, we do not report the detailed results of these initial investigations in the results Section.

- Use more powerful base networks. We experiment with Xception and InceptionV3 models instead of MobileNetV2. These more complex models show worse performance/accuracy tradeoff in our experiments, and as a result are less suitable for our application goals.
- Enable regularization. We set the default non-zero Keras regularization coefficients on the dense layers (the compined L1 & L2 parameter, for both and kernel and



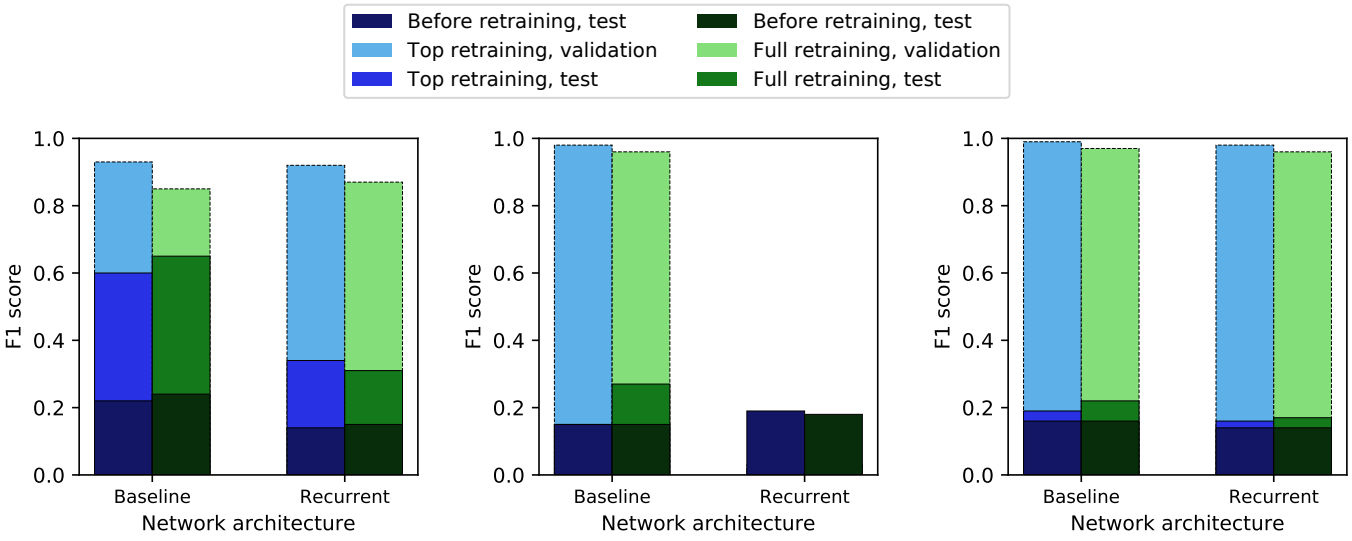
(b) Kaggle dataset (c) METC dataset (d) Hospital dataset

Fig. 6: F_1 scores of the different CNN architectures.



(a) Kaggle dataset (b) METC dataset (c) Hospital dataset

Fig. 7: F_1 scores of the different CNN architectures with two additional dense layers.



(b) Kaggle to METC dataset (c) Kaggle to hospital dataset (d) METC to hospital dataset

Fig. 8: F_1 scores of the different CNN architectures, transfer learning.

activity regularization), and measure generalization accuracy after the change. No improvement is detected.

IV. Results

A. Main Experiments

Fig. 6 shows the main results. Due to size limitations, we only report F_1 scores in this paper. Contrary to the expectations, the best performance on Kaggle and METC datasets is achieved by the baseline model, which only uses a single frame as the input (0.96 F_1 score on Kaggle, in Fig. 6b, 0.64 F_1 score on METC, in Fig. 6c). In particular, full retraining usually improves the accuracy of the baseline classifiers, while decreases it on the more complex classifiers. This suggests that the more complex ones are more likely to overfit.

Most concerningly, none of the approaches show even average performance on the test data on the hospital dataset (Fig. 6d)). The F_1 score of a random classifer is $1/7$ =0.14; the best real classifier shows F_1 score of only 0.21. The poor accuracy is not caused by a failure to learn: the F_1 scores on the validation data is above 0.95 in some of the experiments. The problem is in the poor generalization performance. We further verified this by conducting a separate experiment, which showed that the accuracy is greatly improved the test data consists of a set of videos taken using a camera location that is already present in the training data.

As the two-stream network does not show a major improvement on the two other approaches in any of the experiments, and its characteristics are in-between the other two, we exclude this network from further experiments.

B. Additional Dense Layers

The results (Fig. 7) of these extended classifiers show the same pattern as the resuls in the main experiments. In absolute values, the F_1 scores are a few percent worse on the average, again suggesting that the extra complexity added by the extra layers makes the classifiers more likely to overfit.

C. Transfer Learning

The results (Fig. 8) show that one of the retrained models achieves the best performance on the METC dataset among all experiment groups (0.65 F_1 score) and one on the hospital data (0.27 F_1 score). However, other than this, the lessons learned from attempting to transfer the training between datasets is not encouraging. None of the classifiers show acceptable performance before retraining them on the new dataset. After retraining, the accuracy on the new dataset is on the average similar to the accuracy when the classifier is trained straightly from the MobineNetV2 base. Moreover, the recurrent CNN initially trained on Kaggle data completely failed to learn on the hospital data (Fig. 8c). This suggests the classsifications learned by the models is not transferrable to new situations.

V. Conclusions

We formulate five intuitive hypotheses in the Introduction of this paper, and evaluate them using three different datasets and multiple CNN architectures. Surprisingly, the evaluation shows evidence against all of these hypotheses. Lightweight CNN classifiers that show good results on the Kaggle dataset ($>$0.95 F_1 scores) demonstrate mediocre performance on a more complex lab-based dataset (0.5–0.6 F_1 scores), and fail to generalize on a real-life dataset (H1), despite applying the Keras data augmentation layers to reduce overfitting (H5). Adding temporal information (H2) or more layers (H3) typically reduces the generalization performance, which can be explained by the more complex models being more vulnerable to overfitting. The effect of full retraining (H4) depends on the architecture. These results show that the dataset is in fact more important than the approach when evaluating hand washing movement classification accuracy, and that translating the existing work on hand washing movement classification from the lab to the field is not straightforward. The future work must focus on discovering the root causes for the poor generalization performance.

Acknowledgment

We thank Ansis Skadins for some initial contributions to this work, and Alessandro Masullo for his valuable advice and recommendations for improvements.

References

[1] Kaggle Hand Wash Dataset. https://www.kaggle.com/realtimear/hand-wash-dataset, 2019.

[2] Kevin Cikel, Mario Arzamendia, Derlis Gregor, Daniel Gutierrez, and Sergio Toral. Evaluation of a CNN+ LSTM system for the classification of hand-washing steps.

[3] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[5] Martins Lulla et al. Influence of different types of real-time feedback on hand washing quality assessed with neural networks/simulated neural networks. In *Proceedings of 8th International Multidisciplinary Research Conference SOCIETY. HEALTH. WELFARE*, 2021.

[6] Martins Lulla, Aleksejs Rutkovskis, et al. Hand-washing video dataset annotated according to the world health organization's hand-washing guidelines. *Data*, 6(4):38, 2021.

[7] Akash Nagaraj, Mukund Sood, Chetna Sureka, and Gowri Srinivasa. Real-time Action Recognition for Fine-Grained Actions and The Hand Wash Dataset.

[8] Esa Prakasa and Bambang Sugiarto. Video analysis on handwashing movement for the completeness evaluation. In *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pages 296–301. IEEE, 2020.

[9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv:1406.2199*, 2014.

[10] World Health Organization. *WHO guidelines on hand hygiene in health care: first global patient safety challenge clean care is safer care*. 2009.

[11] Kazushi Yamamoto, Miho Yoshii, Fumiya Kinoshita, and Hideaki Touyama. Classification vs Regression by CNN for Handwashing Skills Evaluations in Nursing Education. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 590–593. IEEE, 2020.

[12] Serena Yeung, Alexandre Alahi, et al. Vision-based hand hygiene monitoring in hospitals. In *AMIA*, 2016.