Hand-Washing Movement Recognition Based on Hand Skeleton Landmarks

Maksims Ivanovs*,†, Ilze Ivanova†

University of Latvia, Faculty of Science and Technology, Raina bulvaris 19, Riga, LV-1586, Latvia

Abstract

Hand hygiene is one of the most effective measures for reducing the spread of healthcare-associated infections, yet adherence to recommended hand-washing protocols is difficult to monitor in clinical environments. Automated monitoring systems based on machine learning and computer vision approaches offer a promising solution by recognizing hand-washing movements from video data, yet approaches relying on raw RGB images are often insufficient. In this work, we investigate the use of skeleton-based representations for recognizing hand-washing movements. Hand keypoints are extracted from video frames using the MediaPipe hand tracking system and represented as temporal sequences describing the spatial configuration of the hands. These sequences are used as input to a sequential neural network that models the temporal dynamics of hand-washing movements. We evaluate the proposed approach on the METC dataset and compare it with a previously proposed RGB-based method. The results show that the skeleton-based representation improves recognition performance, increasing the macro F1 score from 65% to 70.5%. These findings suggest that modeling hand movements using skeletal keypoints provides a promising alternative to raw RGB-based approaches for automated hand-washing movement recognition.

Keywords

Human activity recognition, hand hygiene monitoring, sequential neural networks, LSTM, GRU, 1D CNN

1. Introduction

Proper hand hygiene is one of the most effective measures for preventing healthcare-associated infections [1]. Adherence to recommended hand-washing protocols such as the WHO hand-washing guidelines [1] helps to reduce the transmission of pathogens in clinical environments [2] and is therefore a key component of patient safety. Despite its importance, compliance with hand hygiene guidelines among healthcare professional is often suboptimal [3], [4]. Manual monitoring of hand-washing practices is time-consuming and difficult to scale, which has motivated the development of automated methods for assessing hand hygiene compliance.

Recent advances in machine learning and computer vision have enabled automated analysis of human activities in video recordings [5], [6]. In particular, neural network-based approaches have demonstrated strong performance in various action recognition tasks. These methods are capable of learning complex visual and temporal patterns from data and can therefore be used to recognize the sequential stages of hand-washing procedures. Automated systems based on such methods have the potential to support monitoring and feedback systems in healthcare environments.

Many existing approaches (see e.g. [7] and [8]) rely on RGB video data to recognize hand-washing movements. While such methods can capture relevant visual information, they are also sensitive to variations in lighting conditions, occlusions, and changes in scene composition and camera position. In practice, these factors can significantly degrade the performance of vision-based models in real-world clinical settings. Furthermore, RGB-based models may learn spurious correlations unrelated to the actual hand movements.

Baltic DB&IS 2026 Conference Forum and Doctoral Consortium, 28 June - 1 July 2026, Tartu, Estonia

*Corresponding author.

†These authors contributed equally.

m.ivanovs@lu.lv (M. Ivanovs); ii13004@edu.lu.lv (I. Ivanova)

0000-0003-2477-7327 (M. Ivanovs); 0009-0000-2404-9229 (I. Ivanova)

One of the promising alternative representations is provided by skeletal models that capture the geometric configuration and spatial positions of hand landmarks. Such representations focus directly on the motion and structure of the relevant anatomical components while discarding much of the irrelevant visual information present in raw images. This can potentially make action recognition models more robust to environmental variability.

In this work, we investigate the use of hand landmark sequences extracted with the MediaPipe framework [9] for recognizing hand-washing movements. The study is conducted on the METC dataset [10], which contains annotated recordings of hand-washing sessions. Instead of training neural network models directly on RGB frames, we represent each frame using a set of detected hand landmarks and model their temporal dynamics using recurrent neural networks.

Our hypothesis is that skeletal representations of hand motion can improve accuracy of hand-washing movement classification, as they provide a more informative and robust signal for recognizing hand hygiene stages than raw RGB data. The goal of this study is therefore to evaluate the effectiveness of landmark-based sequential models for automatic hand-washing movement recognition.

2. Background

2.1. Hand washing monitoring in healthcare

Hand hygiene has long been recognized as an important issue in epidemiology. It was discussed particularly widely during the recent COVID-19 pandemic, when contaminated hands were initially believed to be a major route of transmission for the coronavirus SARS-CoV-2. However, even after it became clear that the virus is primarily airborne and the public attention to hand hygiene declined, the issue has remained important for healthcare professionals. One reason is that poor hand hygiene is one of the major pathways for the spread of healthcare-associated infections [11], which result in almost 9 million cases per year in Europe alone [12] and therefore represent a major threat in healthcare settings.

To improve health hygiene, there are well-known and seemingly easy-to-follow protocols such as the WHO guidelines for hand washing [1], which comprise six key hand-washing movements that have to be performed for 40 to 60 seconds overall. However, although the grave impact of healthcare-associated infections and the link between poor hand hygiene habits and these habits are well known to healthcare professionals, the levels of compliance with protocols remain low [4], [13]. A recent study conducted in Latvia [14] found that healthcare workers correctly perform all steps of the WHO hand-washing protocol in a real-world clinical setting only in a minority of cases.

To evaluate and improve hand hygiene practices among healthcare professionals, monitoring of their performance can be useful. The traditional approach relies on monitoring conducted by human observers; however, continuous observation is difficult to maintain in practice. Moreover, monitoring conducted only occasionally may lead to the Hawthorne effect, where improvements occur only while the subjects are being observed but disappear at other times. Therefore, a more promising approach is the development of automated monitoring systems. Such systems should be able to provide real-time evaluation of hand-washing episodes with sufficient accuracy, address the domain shift problem, i.e., perform well in new locations and with new users, and preserve user privacy [15]. Ideally, the system should also operate on low-power hardware such as edge devices, which are cost-efficient and easier to deploy in clinical environments.

2.2. Machine learning methods for recognizing handwashing movements

Considering the nature of the hand-washing movement recognition task, using machine learning methods is a promising approach. In several studies [16] [17], various types of wearable sensors have been used. Although these approaches have achieved good results, the requirement to wear additional devices may complicate the procedure and reduce compliance. Moreover, the surfaces of wearable devices themselves may become contaminated with pathogens and thereby contribute to their spread.

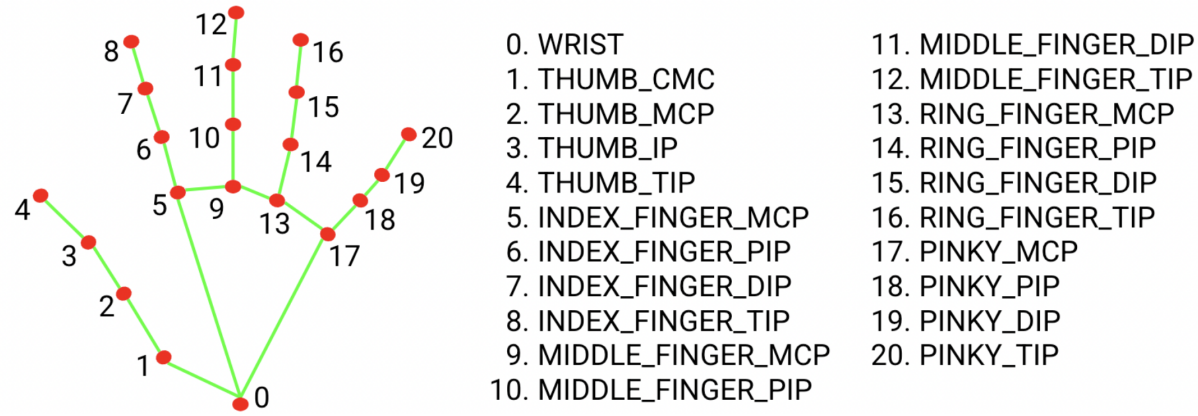


Figure 1: MediaPipe Hands landmarks. Reproduced from Google AI for Developers documentation under the Creative Commons Attribution 4.0 License [31].

Arguably, a more promising approach is to combine machine learning with computer vision methods. Some earlier studies employed pre-deep learning approaches, such as particle filter-based classification [18] and complex pipelines combining feature engineering techniques with an SVM-based classifier [19]. However, most modern computer vision solutions rely on deep neural networks. A typical architecture for gesture recognition consists of a convolutional neural network (CNN) combined with a sequential model such as Gated Recurrent Units (GRU; [20]) or Long Short-Term Memory networks (LSTM; [21]) to capture the temporal dynamics of hand-washing gestures. To further improve the performance of automated monitoring systems, some studies employ depth cameras [22] or multi-camera setups [23]. While these approaches appear promising, they may also increase both the cost and the complexity of deploying such systems.

A possible alternative representation of hand movements in video is provided by hand skeletons. Neural network-based tools such as the MediaPipe Hands framework [9] make it possible to extract hand landmarks from RGB videos with high accuracy. These landmarks can then be used as input features for sequential neural networks such as GRU or LSTM. As the MediaPipe Hands output consists of 21 landmarks per hand (see Fig. 1), its spatial resolution is sufficient for classifying complex gestures. Consequently, it has been used as a feature extractor for tasks such as sign language recognition [24, 25] and keyboard typing gesture classification [26].

In recent studies, hand skeletons extracted with MediaPipe have also been used for hand-washing gesture recognition. We provide a brief overview of some of such studies in the following.

In [27], MediaPipe Hands was used to extract features from RGB-D videos. While the authors report good results, the proposed model requires depth sensors and therefore cannot be directly benchmarked on standard RGB datasets.

In [28], MediaPipe Hands was used together with three classical machine learning algorithms—Logistic Regression, Random Forest, and Support Vector Machines—trained on a custom-collected dataset. The best model achieved a near-perfect accuracy of 99.5%. However, the videos in the dataset appear to feature near-ideal recording conditions and highly accurate gesture execution, which likely contributes to the unusually high performance of the model.

Finally, in [29], MediaPipe Hands and LSTM were applied to the classification of three datasets with varying levels of complexity: the publicly available Kaggle hand-washing dataset [30], METC [10], and Handwash [14]. The authors report very good performance of their model on all three datasets; however, accuracy and F1 scores are reported only for the training and validation splits, while results on unseen test data are not provided. Furthermore, the description of the methodology omits several details essential for reproducibility, such as the number of training epochs and the type of normalization applied to the MediaPipe data. Therefore, although the MediaPipe-based approach was evaluated on more challenging datasets such as METC and Handwash, further research is needed to fully assess its effectiveness.

(a) (b) (c)

Figure 2: Sample images from: (a) the Kaggle hand-washing dataset; (b) the METC dataset; (c) the PSKUS dataset.

2.3. Datasets of handwashing recordings

Since the machine learning paradigm implies that models learn from data, datasets containing recordings of hand-washing movements play a crucial role in the development of automated hand-washing monitoring systems. Many studies rely on custom-collected datasets, which often feature controlled conditions such as good lighting and careful execution of all steps of the WHO hand-washing protocol. These datasets are typically relatively small, often containing only several dozen recordings.

Perhaps the most widely used dataset for experiments on automatic hand-washing evaluation—the publicly available portion of the Kaggle hand-washing dataset—follows a similar pattern, as it contains only 25 hand-washing episodes featuring carefully executed movements. The number of larger publicly available datasets with hand-washing videos is comparatively small. For example, the Kinetics Human Action Video Dataset [32] contains 916 hand-washing videos, while the STAIR Actions dataset [33] includes around 1,000 such videos. However, neither dataset provides labeling according to the WHO protocol, which makes them difficult to use for supervised learning in this task.

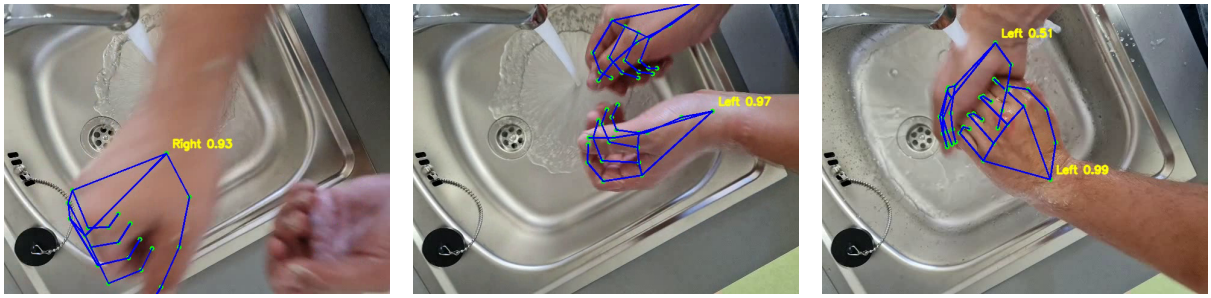
A notable example of a large real-world hand-washing video dataset is Handwash [14], which contains 3,185 labeled videos. However, despite being published five years ago, it has been used only in a limited number of studies. This may be due to its challenging nature, as the dataset includes videos recorded in different locations and featuring imperfect execution of hand-washing techniques.

An important consideration when using datasets for research on hand-washing movement classification is the limited transferability of results obtained on simpler datasets to real-world conditions. In particular, a recent study [34] comparing three datasets of different complexity (see sample images in Fig. 2) demonstrated that near-perfect accuracy achieved by CNN-based models on the simpler Kaggle dataset dropped to a macro F1 score of 65% when the same architectures were trained and evaluated on the more complex METC dataset, and further declined to near-random performance when evaluated on the challenging Handwash dataset. These results indicate that the apparent success of models on controlled datasets does not necessarily translate to robust performance in realistic environments.

3. Methodology

The goal of this study was to evaluate the effectiveness of landmark-based sequential models for automatic hand-washing movement recognition. To this end, we conducted experiments on the METC dataset, which is more complex than laboratory-based datasets such as the Kaggle hand-washing dataset but less complex than the Handwash dataset, and therefore well suited for preliminary experiments.

For the experiments, hand keypoints were extracted from the METC videos using the MediaPipe Hands framework. Several feature representations of the resulting hand skeletons were then considered, including different normalization strategies and alternative skeletal feature constructions. Because skeleton detections were occasionally missing for some frames, we also evaluated the effect of temporal imputation based on interpolation from neighbouring frames. Importantly, normalization methods and imputation were treated as experimental conditions rather than mandatory preprocessing steps.



(a) (b) (c)

Figure 3: Sample hand skeletons produced with MediaPipe Hand shown as overlay over corresponding RGB frames.

Finally, several sequential models, namely, LSTM, GRU, and one-dimensional convolutional neural networks (1D CNN), were trained on the resulting hand skeleton sequences and evaluated. We describe all these aspects of the methodology in more detail in the rest of this section.

3.1. METC dataset

The METC dataset consists of 212 videos accompanied by JSON files containing frame-level annotations of hand-washing movements. The videos were labeled in real time during data acquisition, which may introduce small temporal delays between the actual movement and the recorded label. The videos have a resolution of 640×480 pixels and a frame rate of 16 frames per second. The overall duration of the videos is 13 870 seconds, and the overall duration of hand-washing movements is 9 144 second.

In our experiments, the dataset was divided into training, validation, and test subsets using a 70/10/20 split. The split was performed at the video level to ensure that frames from the same video do not appear in multiple subsets.

3.2. Hand keypoint extraction

Hand keypoints were extracted from the videos using MediaPipe Hands, a real-time hand tracking framework that estimates anatomical landmarks of the hand using a deep neural network. For each detected hand, the system predicts the locations of 21 landmarks corresponding to the wrist and finger joints. Each landmark is represented by three coordinates (x, y, z) describing its position relative to the image. Sample hand skeletons are shown in Fig. 3.

For every frame, the extracted landmark coordinates were stored together with metadata including the video name, frame index, timestamp, presence indicators for the left and right hand, detection confidence scores, and the movement label. If a hand was not detected in a given frame, the presence flag was set to zero and the corresponding landmark coordinates were recorded as zeros.

3.3. Normalization of hand keypoints

To investigate the effect of landmark preprocessing on model performance, we evaluated three alternative normalization configurations: (i) no normalization, (ii) location normalization only, and (iii) location and scale normalization applied jointly. The normalization procedures followed the approach proposed in [35].

Location normalization was performed by subtracting the coordinates of the wrist landmark (landmark 0) from all other landmarks of the same hand:

$$\tilde{x}_i = x_i - x_{wrist}, \quad \tilde{y}_i = y_i - y_{wrist}, \quad \tilde{z}_i = z_i - z_{wrist}.$$

This transformation expresses all landmarks relative to the wrist position, thereby removing the effect of the global hand location in the frame.

In the third configuration, an additional scale normalization was applied by dividing all landmark coordinates in a frame by the maximum absolute coordinate magnitude observed in that frame:

$$x_i' = \frac{\tilde{x}_i}{s}, \qquad s = \max(|\tilde{x}_i|, |\tilde{y}_i|, |\tilde{z}_i|).$$

This step reduces variability caused by differences in hand size or distance from the camera.

In addition to the raw landmark coordinates, alternative feature representations were also derived from the landmarks. These included bone vectors describing the displacement between connected joints and pairwise distances between selected landmark pairs, such as distances between the wrist and fingertips and between fingertip pairs.

3.4. Handling missing detections

Occasionally, MediaPipe fails to detect a hand in a frame. In the original annotation files, such cases are encoded by setting the presence flag of the corresponding hand to zero and storing all landmark coordinates as zero.

To investigate the effect of missing detections on model performance, we considered two experimental conditions: (i) using the raw landmark sequences without modification and (ii) applying interpolation to fill missing detections.

In the interpolation condition, missing landmark coordinates were handled using linear interpolation. For each hand independently, landmark coordinates corresponding to frames with missing detections were first converted to missing values. Linear interpolation was then applied along the temporal dimension to fill gaps occurring between valid detections.

Interpolated frames were marked using imputation flags for the left and right hand. If missing detections occurred at the beginning or end of a sequence, where interpolation was not possible, the landmark coordinates were set to zero and the presence flag remained unset.

3.5. Sequence construction

Handwashing movements are inherently temporal and cannot be reliably recognised from individual frames. To capture temporal structure, the landmark features were organized into short sequences using a sliding window approach. Each sequence consisted of a fixed number of consecutive frames and represented a short segment of hand motion.

For the LSTM experiments, two window lengths were considered: 15 and 30 frames. For the GRU and temporal CNN experiments, the window lengths were 24 and 48 frames. In all cases, windows were generated with a stride of one frame. Each sequence was assigned the label corresponding to the final frame in the window. Windows were constructed only within contiguous frame segments to ensure that sequences did not span gaps in frame indices.

Only windows in which at least one hand was detected in every frame were retained. Sequences were not required to have a constant label throughout the entire window. No transition trimming was applied.

Before sequence construction, feature values were standardized using a `StandardScaler` fitted on all frames of the training set. The same transformation was then applied to validation and test data.

3.6. Models and training procedure

Three types of temporal models were evaluated: long short-term memory networks (LSTM), gated recurrent units (GRU), and one-dimensional convolutional neural networks (1D CNN).

LSTM. Long short-term memory networks are a type of recurrent neural network designed to capture temporal dependencies in sequential data. LSTM units maintain an internal memory state controlled by input, forget, and output gates, allowing the network to retain or discard information over time. In this study, the LSTM network consisted of two or three stacked recurrent layers with hidden state dimensionality of either 64 or 128 units. The final hidden state was used as a sequence representation

Table 1
LSTM architecture configurations evaluated in the experiments

Parameter	Values
Hidden state size	64, 128
Number of LSTM layers	2, 3
Dropout	0.2
Sequence length	15, 30 frames
Stride	1
Classifier	Fully connected layer (7 classes)

Table 2
GRU architecture configurations evaluated in the experiments

Parameter	Values
Hidden state size	128, 256
Number of GRU layers	1, 2
Dropout	0.2
Sequence length	24, 48 frames
Stride	1
Classifier head	Linear(128) + ReLU + Dropout + Linear(7)

Table 3
1D CNN architecture configurations evaluated in the experiments

Parameter	Values
Number of convolutional layers	2, 3
Number of filters	64, 128
Kernel size	3, 5
Activation function	ReLU
Pooling	Max pooling
Classifier	Fully connected layer (7 classes)

and passed through a dropout layer with a rate of 0.2 followed by a fully connected layer producing class logits for the seven movement classes. The summary of the LSTM architectures is given in Table 1.

GRU. Gated recurrent units are a simplified variant of LSTM networks. GRUs combine the input and forget gates into a single update gate and use a reset gate to control the influence of previous hidden states, reducing the number of parameters while preserving the ability to model temporal dependencies. In this study, the GRU network consisted of one or two recurrent layers with hidden state dimensionality of either 128 or 256 units. The final hidden state was passed to a multilayer perceptron classification head comprising a linear layer with 128 units, a ReLU activation function, a dropout layer with a rate of 0.2, and a final linear layer producing class logits for the seven movement classes. The summary of the GRU architectures is given in Table 2.

1D CNN. One-dimensional convolutional neural networks model temporal patterns using convolutional filters applied along the time dimension. In this study, the temporal CNN received input in the form of feature sequences and applied one-dimensional convolutions after transposing the input so that the feature dimension became the channel dimension. The network consisted of two convolutional layers with either 64 or 128 filters and kernel sizes of 3 or 5. Each convolutional layer was followed by a ReLU activation function and a dropout layer with a rate of 0.2. The extracted temporal features were aggregated using adaptive average pooling, after which they were passed to a classification head consisting of a linear layer with 64 hidden units, a ReLU activation function, a dropout layer, and a final linear layer producing class logits for the seven movement classes. The summary of the 1D CNN architectures is given in Table 3.

Table 4
Best performance achieved by each model type.

Model	Test Accuracy	Test Macro F1
LSTM	65.9%	65.7%
GRU	69.7%	68.1%
Temporal CNN	70.3%	70.5%

All models were implemented in PyTorch and trained on an NVIDIA A100 GPU. Training was performed for 100 epochs using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 256. During training, validation performance was monitored and the model checkpoint with the highest validation macro F1 score was selected. After training, the selected checkpoint was evaluated on the held-out test set.

For the LSTM, GRU, and temporal CNN models, experiments were performed for all combinations of feature representations, imputation settings, model hyperparameters, and sequence lengths. This resulted in 96 experimental configurations for each model architecture.

4. Results and discussion

This section presents the experimental results obtained with the evaluated sequential neural network models. Performance was evaluated using macro F1 score and classification accuracy on the held-out test set.

4.1. LSTM results

The best-performing LSTM configuration used the raw landmark representation with imputation enabled, a hidden state size of 64, two recurrent layers, and a window size of 30 frames. This configuration achieved a test macro F1 score of 65.7% and a test accuracy of 65.9%.

Across LSTM experiments, performance differences between feature representations were relatively small. Window sizes of 30 frames generally produced slightly better results than shorter sequences.

4.2. GRU Results

The GRU models consistently outperformed the LSTM models. The best GRU configuration used the `norm_zoom` representation without imputation, with a hidden size of 128, two recurrent layers, and a window size of 48 frames. This configuration achieved a test macro F1 score of 68.1% and a test accuracy of 69.7%.

The results suggest that GRU architectures are able to capture temporal dependencies in the hand landmark sequences more effectively than LSTM models while using fewer parameters.

4.3. Temporal CNN results

The best overall performance in the study was achieved by the temporal CNN model. The highest test macro F1 score of 70.5% was obtained using the `norm_loc` representation with imputation enabled, 64 convolutional filters, kernel size 3, and a window size of 48 frames.

In general, longer temporal windows produced better results for the convolutional models. The temporal CNN models also showed relatively stable performance across different feature representations.

4.4. Overall model comparison

Table 4 summarises the best results obtained with each model type.

The temporal CNN achieved the best overall performance, reaching a macro F1 score of 70.5% on the test set. The GRU model achieved the second-best result with a macro F1 score of 68.1%, while the LSTM model achieved a maximum macro F1 score of 65.7%.

These results indicate that convolutional temporal modelling is well suited for the recognition of handwashing movements represented by skeletal keypoints. Furthermore, compared with the results on the METC dataset in [34], the proposed MediaPipe-based approach improves the macro F1 score from 65% obtained in previous experiments based on RGB representations to $\approx$ 70% obtained with the temporal CNN model in this study. This improvement suggests that skeletal representations extracted from hand landmarks provide a more informative and compact description of hand-washing movements than raw image features for this task.

5. Conclusion and future work

In this study, we investigated the use of hand skeleton representations extracted with the MediaPipe Hands framework for recognizing hand-washing movements. Instead of learning directly from raw RGB frames, the proposed approach models the temporal dynamics of hand landmark sequences using sequential neural network architectures. The method was evaluated on the METC dataset, which represents a moderately challenging benchmark for hand-washing movement recognition.

The experimental results demonstrate that skeletal representations provide a viable alternative to RGB-based approaches. Among the evaluated architectures, the temporal CNN model achieved the best performance, reaching a macro F1 score of 70.5% on the test set, followed by the GRU model with 68.1% and the LSTM model with 65.7%. Compared with previously reported RGB-based results on the same dataset, the proposed approach improves the macro F1 score from 65% to approximately 70%. These findings support the hypothesis that modeling hand movements using skeletal keypoints can provide a more informative and compact representation of hand-washing gestures while reducing the influence of irrelevant visual information present in raw images.

The results also indicate that convolutional temporal modeling is particularly well suited for analyzing sequences of hand landmarks. At the same time, the overall performance levels suggest that recognizing complex hand-washing movements in realistic conditions remains a challenging problem.

Several directions for future work can be identified. First, the proposed approach should be evaluated on more challenging datasets such as the Handwash dataset to better assess its robustness in real-world clinical environments. Second, additional feature representations derived from hand landmarks could be explored, including joint angles, relative finger orientations, and graph-based skeletal representations. Third, more advanced temporal modeling approaches, such as transformer-based architectures or hybrid convolutional–recurrent models, may further improve recognition performance. Fourth, future studies could include repeated experiments with multiple random seeds or cross-validation procedures to obtain more robust estimates of model performance, as the present study relied on a single train/validation/test split of the dataset.

Another promising direction is the integration of skeletal representations with complementary visual cues. For example, combining hand landmarks with coarse RGB features or depth information could potentially improve robustness in cases where hand tracking becomes unreliable. Finally, future studies should investigate the feasibility of deploying such models on edge devices for real-time hand hygiene monitoring systems in healthcare settings.

Overall, the findings of this study demonstrate that hand skeleton representations extracted with MediaPipe provide a promising foundation for automated hand-washing movement recognition and may contribute to the development of practical monitoring systems that support hand hygiene compliance in clinical environments.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-5.5 in order to assist with LaTeX formatting. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] World Health Organization, WHO guidelines on hand hygiene in health care, World Health Organization, 2009.

[2] K. J. Mckay, R. Z. Shaban, P. Ferguson, Hand hygiene compliance monitoring: Do video-based technologies offer opportunities for the future?, Infection, Disease & Health 25 (2020) 92–100.

[3] V. Erasmus, T. J. Daha, H. Brug, J. H. Richardus, M. D. Behrendt, M. C. Vos, E. F. Van Beeck, Systematic review of studies on compliance with hand hygiene guidelines in hospital care, Infection Control & Hospital Epidemiology 31 (2010) 283–294.

[4] D. J. Gould, D. Moralejo, N. Drey, J. H. Chudleigh, M. Taljaard, Interventions to improve hand hygiene compliance in patient care, Cochrane database of systematic reviews (2017).

[5] G. Saleem, U. I. Bajwa, R. H. Raza, Toward human activity recognition: a survey, Neural Computing and Applications 35 (2023) 4145–4182.

[6] M. Karim, S. Khalid, A. Aleryani, J. Khan, I. Ullah, Z. Ali, Human action recognition systems: A review of the trends and state-of-the-art, IEEE Access 12 (2024) 36372–36390.

[7] T. Xie, J. Tian, L. Ma, A vision-based hand hygiene monitoring approach using self-attention convolutional neural network, Biomedical Signal Processing and Control 76 (2022) 103651.

[8] C. Zhong, A. R. Reibman, H. A. Mina, A. J. Deering, Designing a computer-vision application: A case study for hand-hygiene assessment in an open-room environment, Journal of Imaging 7 (2021) 170.

[9] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, M. Grundmann, Mediapipe: A framework for building perception pipelines, 2019. URL: https://arxiv.org/abs/1906.08172. arXiv:1906.08172.

[10] O. Zemlanuhina, et al., Influence of different types of real-time feedback on hand washing quality assessed with neural networks/simulated neural networks, in: SHS Web of Conferences, volume 131, EDP Sciences, 2022, pp. 1–13.

[11] P. Bolton, T. J. McCulloch, The evidence supporting who recommendations on the promotion of hand hygiene: a critique, BMC Research Notes 11 (2018) 899.

[12] C. Suetens, et al., Prevalence of healthcare-associated infections, estimated incidence and composite antimicrobial resistance index in acute care hospitals and long-term care facilities: Results from two European point prevalence surveys, 2016 to 2017, Eurosurveillance 23 (2018) 1–17.

[13] V. Mouajou, K. Adams, G. DeLisle, C. Quach, Hand hygiene compliance in the prevention of hospital-acquired infections: a systematic review, Journal of Hospital Infection 119 (2022) 33–48.

[14] M. Lulla, A. Rutkovskis, A. Slavinska, A. Vilde, A. Gromova, M. Ivanovs, A. Skadins, R. Kadikis, A. Elsts, Hand-washing video dataset annotated according to the world health organization's hand-washing guidelines, Data 6 (2021) 38.

[15] M. Ivanovs, Deep Learning for Applied Computer Vision: Solving Image Understanding Tasks with Convolutional Neural Networks, Phd thesis, University of Latvia, Riga, Latvia, 2024.

[16] Y. Cao, F. Li, H. Chen, X. Liu, S. Yang, Y. Wang, Leveraging wearables for assisting the elderly with dementia in handwashing, IEEE Transactions on Mobile Computing 22 (2022) 6554–6570.

[17] C. Wang, et al., Accurate measurement of handwash quality using sensor armbands: Instrument validation study, JMIR MHealth Uhealth 8 (2020) e17001.

[18] J. Hoey, A. Von Bertoldi, P. Poupart, A. Mihailidis, Assisting persons with dementia during handwashing using a partially observable Markov decision process., in: Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007), 2007.

[19] D. F. Llorca, I. Parra, M. Á. Sotelo, G. Lacey, A vision-based system for automatic hand washing quality assessment, Machine Vision and Applications 22 (2011) 219–234.
[20] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, arXiv:1409.1259 (2014).
[21] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
[22] A. Greco, G. Percannella, P. Ritrovato, A. Saggese, M. Vento, A deep learning based system for handwashing procedure evaluation, Neural Computing and Applications 35 (2023) 15981–15996.
[23] H. Ren, G. Lin, W. Lian, F. Jing, Y. Zou, Y. Liu, Q. Zhang, K. Wu, W. Cheng, Automated monitoring of hand hygiene compliance using multi-camera systems in healthcare environments, IEEE Internet of Things Journal (2025).
[24] R. Kumar, A. Bajpai, A. Sinha, Mediapipe and cnns for real-time asl gesture recognition, arXiv preprint arXiv:2305.05296 (2023).
[25] A. R. Verma, G. Singh, K. Meghwal, B. Ramji, P. K. Dadheech, Enhancing sign language detection through mediapipe and convolutional neural networks (cnn), arXiv preprint arXiv:2406.03729 (2024).
[26] B. Mallik, M. A. Rahim, A. S. M. Miah, et al., Virtual keyboard: A real-time hand gesture recognition-based character input system using lstm and mediapipe holistic., Computer Systems Science & Engineering 48 (2024).
[27] Y. Zhang, T. Maekawa, Interhandnet: Capturing two-hand interaction for robust hand-washing activity recognition, in: 2025 IEEE International Conference on Pervasive Computing and Communications (PerCom), IEEE, 2025, pp. 13–24.
[28] J. N. Quiñones-Romero, A. F. Romero-Gómez, R. Buitrago, Development of a system for monitoring and validation of proper hand washing using machine learning, Clinical Epidemiology and Global Health 33 (2025) 101971.
[29] T.-C. Lin, F.-C. Lin, A hybrid deep learning approach for recognizing hand washing steps via skeleton features, in: 2025 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), IEEE, 2025, pp. 535–536.
[30] Sample: Kaggle Hand Wash Dataset, [Online], 2019. Available: https://www.kaggle.com/realtimear/hand-wash-dataset. Accessed 14 February 2026.
[31] Google AI for Developers, Hand landmarks detection guide, https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker, 2026. Last updated: 21 April 2026; accessed: 18 May 2026.
[32] W. Kay, et al., The Kinetics human action video dataset, arXiv:1705.06950 (2017).
[33] Y. Yoshikawa, J. Lin, A. Takeuchi, STAIR actions: A video dataset of everyday home actions, arXiv:1804.04326 (2018).
[34] A. Elsts, M. Ivanovs, R. Kadikis, O. Sabelnikovs, CNN for hand washing movement classification: What matters more–the approach or the dataset?, in: 2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, 2022, pp. 1–6.
[35] M. Á. Remiro, M. Gil-Martín, R. San-Segundo, Improving hand pose recognition using localization and zoom normalizations over mediapipe landmarks, Engineering Proceedings 58 (2023) 69.