Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33389

Impact of Dataset Characteristics on Lightweight CNN-Based Hand Gesture Recognition Using Edge Impulse

Zheng Liao

Department of Computer and Science, Guangzhou University, Guangzhou, China
32306300021@e.gzhu.edu.cn

Abstract. This study analyzes the influence of dataset characteristics on the recognition of hand gestures using a lightweight convolutional neural network (CNN) based on the Edge Impulse platform. Under a fixed CNN architecture, training strategy, and sample size, three public datasets with varying levels of preprocessing quality, background complexity, and inter-class feature similarity were tested in a controlled experimental setup. The results indicate that dataset characteristics have a significant impact on classification performance. The model achieved 95.0% accuracy and a weighted F1-score of 0.95 for a well-preprocessed dataset with clear feature separability. In contrast, performance dropped to 35.0% correctly classified samples on the dataset with complex backgrounds. On the third dataset, an 86.3% accuracy was obtained; however, performance was class-dependent due to high inter-class feature similarity. In summary, the findings verify that preprocessing quality, background complexity, and inter-class feature dissimilarity are the key factors determining the performance of lightweight gesture recognition in Edge Impulse-based CNNs.

Keywords: Hand Gesture Recognition, Convolutional Neural Networks, Dataset Characteristics, Edge AI

1. Introduction

With the growth of the Internet of Things (IoT), wearable devices, and embedded systems, ubiquitous computing has shifted from a theoretical concept to widespread industrial practice, and the core paradigm of human-computer interaction is accelerating its transition from device-centric to human-centric natural interaction [1]. Meanwhile, traditional cloud-based Artificial Intelligence (AI) architectures face growing pain points, including high transmission latency, high bandwidth costs, private data leakage risks, and offline scenario inoperability [2]. Therefore, edge intelligence has become a core development direction for both industry and academia, which offloads AI model training and inference capabilities to end devices for local data processing and real-time response [3,4].

Gesture recognition, is a non-contact natural interaction method in line with human intuitive interaction [5,6]. And it has become the core interaction portal for edge intelligent devices with the advantages of intuitiveness, hands-free operation and low intrusiveness, and delivers irreplaceable application value in scenarios including hearing-impaired assisted communication and wearable device shortcut triggering [7].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33389

Hand gesture recognition models' core metrics including recognition accuracy, generalization, and edge inference efficiency are highly dependent on training dataset characteristics. Academia has released multiple mainstream public hand gesture datasets, with significant heterogeneity in acquisition modality (vision/IMU/sEMG), sampling parameters, gesture categories, subject scale, scene interference, annotation rules, and sample size, as well as clear gaps in applicable scenarios and optimization targets [8].

Nevertheless, most of the above studies concern either solely the model optimization for a single dataset to improve accuracy or the comparison of the performance of different models on one dataset. Still, a meaningful analysis of training approaches and edge runtime performance (inference latency, resource usage, robustness etc.) across prevalent gesture datasets under one unified model, the Edge Impulse toolchain, and the hardware deployment platform is missing [9]. This is the objective that this work intends to achieve in the domain of hand gesture recognition since this analysis is essential to the choice of datasets, the adaptation of models, and the optimization of deployment for edge-side applications.

2. Method

This work collects 3 different common visual datasets from Kaggle. There are significant differences among the three datasets in terms of image acquisition methods, background complexity, the number of gesture categories and data scale, which can be well used to analyze the impact of dataset characteristics on the performance of gesture recognition models [8].

The first dataset includes 10 gesture classes, each containing around 500 images. The images have had their backgrounds removed using OpenCV, resulting in cleaner data with reduced environmental noise. The second dataset contains approximately 2000 images across 8 gesture categories. Images were captured in real-world environments with complex backgrounds, and the original resolution is 207 × 207 pixels. The third dataset consists of 24,000 images spanning 20 gesture categories, with 900 images per category for training and 300 for testing. The dataset provides rich diversity and a balanced distribution of samples.

2.1. Dataset preparation

This work applied a series of preprocessing steps to make sure a fair comparison across the three datasets shown in Table 1. This work selected five gesture categories from each dataset to maintain consistency across experiments. And the number of images per category was limited to 200, with random sampling applied if a category contained more than 200 images. All images were standardized to 96 × 96 pixels using Edge Impulse's built-in image processing tools to ensure uniform input dimensions for the model [7].

These preprocessing steps prepared the datasets for direct upload to the Edge Impulse platform and ensured that any differences in model performance were primarily caused by dataset characteristics rather than inconsistencies in category count, sample size, or image resolution.

Table 1. The composition of dataset

Dataset ID	Class 1	Class 2	Class 3	Class 4	Class 5
Dataset 1	peace	rock_on	scissor	thumbs	up
Dataset 2	closedFist	fingerSymbols	semiOpenFist	semiOpenPalm	singleFingerBend
Dataset 3	Number_Five	Number_Six	Number_Seven	Number_Eight	Number_Nine

DOI: 10.54254/2755-2721/2026.GU33389

The Figure 1 shows that the first image presents the peace gesture from Dataset 1, the middle image presents the closedFist gesture from Dataset 2, and the final image presents the Number_Five gesture from Dataset 3.



Figure 1. Representative gesture samples from three datasets (picture credit: original)

2.2. Edge impulse-based CNN model

2.2.1. Introduction to convolutional neural networks (CNNs)

CNNs are a deep learning model that has proven to be very effective for image and video recognition. The principle of CNNs is to automatically learn hierarchical features from input images via the cascaded convolution, down-sampling and fully connect layers [10]. Convolutional layers perform local spatial pattern extraction (e.g. edges, textures, shapes) by applying a number of learnable filters; pooling layers achieve this by up-sampling spatial dimension and increase the robustness of features; finally, fully connected layers aggregate these extracted features for classification or regression tasks [11]. CNNs are well-suited for learning spatial hierarchies, require little manual feature engineering, and demonstrate superb performance on visual recognition tasks. Drawbacks include high computational cost, large model size, and susceptibility to overfitting when trained on small datasets [7,9].

2.2.2. Introduction to edge impulse

Edge Impulse is a cloud based platform for developing machine learning models for embedded and edge devices. It offers the entire pipeline from data collection, preprocessing, model training, deployment to inference [9]. Edge Impulse supports image, audio, and sensor data types and users can train models with minimal programming knowledge, which is ideal for quickly prototyping embedded AI applications.

Edge Impulse also provides multiple interfaces for model interaction, including a web-based dashboard for training and monitoring metrics, the Impulse API for uploading new data or triggering inference, and deployment options such as C++ SDK, Python SDK, or TensorFlow Lite for microcontrollers.

2.2.3. Implementation details

This work built a CNN-based image classification model on Edge Impulse for hand gesture recognition. The process included these steps:

Data Upload and Preprocessing: This study uploaded the preprocessed datasets to the Edge Impulse platform. The images were subsequently standardized to 96 × 96 pixels using the platform's built-in image processing block, and the "Image" processing block was utilized to normalize pixel values and reduce color depth when necessary.

DOI: 10.54254/2755-2721/2026.GU33389

Model Training: A CNN learning block was added, and transfer learning was used to fine-tune a pre-trained model on the hand gesture datasets. Key hyperparameters included Training cycles: 30, Learning rate: 0.001, Batch size: 32, Validation set size: 20%, and Training processor: CPU.

2.2.4. Evaluation metrics

The trained CNN models were evaluated with three metrics: Validation Accuracy (validation subset accuracy for overfitting monitoring), Inference Latency (single image processing time on edge hardware), and Resource Usage (RAM and Flash consumption during model deployment), enabling comprehensive evaluation of both model performance and edge deployment suitability.

3. Results and discussion

3.1. Results

The model was trained for 30 epochs with a learning rate of 0.001 and a batch size of 32 on a CPU. The dataset was split into 80% training and 20% validation, without data augmentation, learned optimizer, or class weighting. The final evaluation was performed on the INT8 quantized model. The results are shown in Table 2, Table 3 and Table 4.

For the first dataset, consisting of five hand gesture categories ,the CNN model was trained on Edge Impulse with the same INT8 quantized architecture as used in other experiments. The model achieved excellent overall classification performance. The overall validation accuracy reached 95.0%, with a validation loss of 0.12. The ROC AUC was 1.00, and the weighted precision, recall, and F1-score all reached 0.95, demonstrating strong generalization and stability.

Per-class performance was also high. Only a few misclassifications were observed: PEACE was occasionally misclassified as SCISSOR or THUMBS; ROCK_ON and THUMBS were occasionally misclassified as UP; UP was occasionally misclassified as SCISSOR; and SCISSOR was occasionally misclassified as ROCK_ON. Feature space visualization showed clear cluster separation among the five gesture categories, with minimal overlap, indicating effective feature learning for each gesture.

Table 2. Confusion matrix of dataset 1

	PEACE	ROCK_ON	SCISSOR	THUMBS	UP
PEACE	93.1%	0%	3.4%	3.4%	0%
ROCK_ON	0%	94.1%	0%	0%	5.9%
SCISSOR	0%	3.3%	96.7%	0%	0%
THUMBS	0%	0%	0%	96.8%	3.2%
UP	0%	0%	5.7%	0%	94.3%
F1 SCORE	0.96	0.96	0.94	0.97	0.93

For the second dataset, the CNN model was trained under the same settings as the first dataset. However, the overall classification performance was poor. The model achieved only 35.0% validation accuracy with a high validation loss of 4.00. The ROC AUC was 0.75, and the weighted precision, recall, and F1-score were 0.60, 0.35, and 0.31, respectively, indicating severe insufficiency in feature learning and generalization.

DOI: 10.54254/2755-2721/2026.GU33389

Per-class performance showed significant variation and strong classification bias. FINGERSYMBOL was the only class with relatively better recognition (recall 77.4%, F1=0.38), with few samples misclassified as SEMIOPENFIST or SEMIOPENPALM. SEMIOPENPALM achieved recall 48.3%, F1=0.46, but over 40% of samples were misclassified as FINGERSYMBOL. SEMIOPENFIST recall was only 28.2%, F1=0.35, with over 50% misclassified as FINGERSYMBOL and nearly 20% misclassified as SEMIOPENPALM. CLOSEDFIST recall was 25%, F1=0.39, and the majority of samples were misclassified into FINGERSYMBOL, SEMIOPENPALM, or SEMIOPENFIST. SINGLEFINGERBEND recognition completely failed, with recall 2.7% and F1=0.05, and over 80% of samples misclassified as FINGERSYMBOL.

Table 3. Confusion matrix of dataset 2

	closedFist	fingerSymbols	semiOpenFist	semiOpenPalm	singleFingerBend
closedFist	25%	37.5%	12.5%	25%	0%
fingerSymbols	0%	77.4%	9.7%	12.9%	0%
semiOpenFist	0%	53.8%	28.2%	17.9%	0%
semiOpenPalm	0%	41.4%	10.3%	48.3%	0%
singleFingerBend	2.7%	81.1%	10.8%	2.7%	2.7%
F1 SCORE	0.39	0.38	0.35	0.46	0.05

For the third dataset, consisting of five numeric hand gestures , the CNN model was trained on Edge Impulse with the same INT8 quantized architecture. The model achieved good overall classification performance, with minor misclassifications in certain categories.

The validation accuracy reached as high as 86.3% with a validation loss of 0.34. The ROC AUC was 1.00, and the weighted precision, recall, and F1-score were 0.90, 0.86, and 0.86, respectively, demonstrating that the baseline model had sufficiently strong feature learning and generalization ability to be directly applied to basic numeric gesture classification problems.

Per-class results exhibited a clear two-level structure. NUMBER_EIGHT, NUMBER_FIVE, and NUMBER_SIX obtained 100% recall with no misclassified samples. They achieved F1-scores of 0.87 (NUMBER_EIGHT), 0.82 (NUMBER_FIVE), and 1.00 (NUMBER_SIX). In contrast, NUMBER_SEVEN had a recall of 75% (F1=0.86), and misclassified 12.5% of samples as NUMBER_EIGHT or NUMBER_FIVE. NUMBER_NINE was the worst class with 62.2% recall and F1=0.77; 13.5% of the samples were misclassified as NUMBER_EIGHT and 24.3% as NUMBER_FIVE. These results suggest that the model confused NUMBER_NINE and NUMBER_SEVEN with NUMBER_EIGHT and NUMBER_FIVE, which is a key target for future optimization.

Table 3. Confusion matrix of dataset 2

	Number_Eight	Number_Five	Number_Nine	Number_Seven	Number_Six
Number_Eight	100%	0%	0%	0%	0%
Number_Five	0%	100%	0%	0%	0%
Number_Nine	13.5%	24.3%	62.2%	0%	0%
Number_Seven	12.5%	12.5%	0%	75%	0%
Number_Six	0%	0%	0%	0%	100%
F1 SCORE	0.87	0.82	0.77	0.86	1.00

DOI: 10.54254/2755-2721/2026.GU33389

All three experiments were deployed on the same target hardware (Cortex-M4F 80MHz) using the EON™ Compiler with RAM optimization, and the inference performance of the INT8 quantized model was identical across datasets regardless of differences in classification accuracy or dataset characteristics, with a single-frame inference latency of 1189 ms, peak RAM usage of 232.9 KB, and Flash storage usage of 546.3 KB.

3.2. Discussion

3.2.1. Impact of dataset characteristics on classification performance

The core finding of this experiment is that, under the condition where the model architecture, training strategy, number of categories, and sample size are all fixed, the quality of data preprocessing, background complexity, and inter-class feature differences are the key factors determining the classification performance of the model. The performance differences among the three types of datasets are essentially the direct result of the inherent characteristics of the data itself. The experiment completely excluded the interference of irrelevant variables, and the conclusion has reliable causal support.

For the first type of preprocessed gesture dataset, the model achieved an overall validation accuracy of 95.0%, a weighted average F1 score of 0.95, and the recall rate for each category exceeded 93%. There were only a few misclassified samples, and the visualization results of the feature space also showed that the clustering boundaries for each category were clear and there was no significant overlap, demonstrating strong feature learning ability and generalization performance. The core reason for this excellent performance lies in two points: First, this dataset completed background removal preprocessing through OpenCV, effectively eliminating the interference of environmental noise and complex backgrounds on the core gesture features, allowing the lightweight CNN model to focus on the learning of core distinguishing features such as hand contours and finger shapes; Second, the 5 types of gestures in this dataset have significant inter-class feature differences, and the hand shapes and finger opening/closing states of different gestures have strong distinguishability, significantly reducing the difficulty of feature learning for the model. This result verifies that high-quality, low-noise preprocessed data, combined with clearly distinguishable gesture categories, is the core prerequisite for resource-constrained edge-end lightweight models to achieve high-precision gesture recognition, and also provides a clear optimization direction for the dataset construction of edge-end gesture recognition systems.

The experimental results of the third type of digital gesture dataset further confirm the direct impact of inter-class feature differences on the classification performance of the model. The overall validation accuracy of the model in this dataset reached 86.3%, and the weighted average F1 score was 0.86. The overall classification performance was good, and it also presented a significant polarized feature. The hand contours and finger shapes of the three gestures of Number_Eight, Number_Five, and Number_Six had extremely high distinctions. The model achieved a 100% recall rate with no misclassified samples. However, the gestures of Number_Nine and Number_Seven had a relatively high feature similarity with other categories, and the core distinguishing features of the hand shape were not obvious. Therefore, there were concentrated misclassification situations, becoming the core weakness of the model performance. This also corresponded exactly to the result in the feature space visualization that the two types of samples had feature overlaps with other categories. Under the condition that the sample size was the same as the other two datasets, this result indicates that the distinction of inter-class features directly determines the single-class classification effect of the model. Even if the overall dataset has a good generalization foundation,

DOI: 10.54254/2755-2721/2026.GU33389

the categories with high inter-class feature similarity will still become the performance bottleneck of the model.

3.2.2. Limitations and future work

This study identified the core impacts of preprocessing quality, background complexity, and inter-class feature difference on the classification performance of lightweight gesture recognition models through rigorous controlled experiments. However, several limitations remain. First, only a single CNN architecture and transfer learning fine-tuning strategy were adopted, without comparing the classification performance of other lightweight architectures (e.g., MobileNet, EfficientNet-Lite) on datasets with varying background complexity and inter-class feature similarity; thus, the adaptability differences of different models to complex data remain unclear [12,13]. Second, regarding the classification failure on the complex background dataset (Dataset 2), this study did not further explore the performance improvement effects of optimization strategies including background removal, multi-scale data augmentation, and hard example mining, failing to clarify the optimal optimization pipeline for such complex-scenario datasets. Third, experiments were only conducted on single-frame static gesture classification, without investigating model performance in real-world scenarios such as dynamic gesture sequences, complex illumination, and occlusion, leaving the research scenario coverage to be further expanded.

To overcome these limitations, future work will be focused on the areas as follows. First, for gesture datasets with complex backgrounds, future work will systematically study the influences of preprocessing and training regimens (background removal, multi-scale data augmentation, class balance weighting etc.) on the classification accuracy of lightweight models, to find out the optimal scheme for training gesture recognition models in complex scenarios. Second, future work will conduct a performance comparison among different lightweight architectures on datasets with various characteristics, to screen the optimal model that balances high recognition accuracy, strong generalization ability, excellent inference efficiency, and complex-scenario adaptability. Third, future work will extend the experimental scenarios, conduct model optimization and validation for practical applications including real dynamic gesture recognition and real-time gesture classification under complex illumination with partial occlusion, so as to further boost the engineering application value of the conclusions.

4. Conclusion

This study adopted a rigorous controlled variable method, with fixed model architecture, training parameters, dataset class count and sample size. Through comparative experiments on three gesture datasets, this study clarified the influence of intrinsic dataset characteristics on the gesture classification performance of lightweight CNN models.

Experiments verified that dataset preprocessing quality, background complexity and inter-class feature difference are the core determinants of model classification performance. The model reached 95.0% classification accuracy on well-preprocessed datasets with clear inter-class discriminability, while it suffered classification failure with only 35.0% accuracy on unprocessed complex-background datasets with highly similar inter-class features. Inter-class feature discriminability also directly dominates per-class classification performance.

In conclusion, this study confirms via rigorous controlled comparative experiments that preprocessing quality, background complexity and inter-class feature difference are the core factors dominating the gesture classification performance of lightweight CNN models. With identical model

DOI: 10.54254/2755-2721/2026.GU33389

architecture, training strategy, sample size and class count, the intrinsic characteristics of data directly determine the model's final classification performance. This study provides experimental and theoretical support for dataset construction and optimization in gesture recognition and similar lightweight image classification tasks.

References

[1] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv. https: //arxiv.org/abs/1704.04861
[2] LeCun, Y., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.
[3] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.
[4] Murthy, G. R. S., & Jadon, R. S. (2009). Vision-based hand gesture recognition: A review. Multimedia Tools and Applications.
[5] Nagi, J., et al. (2021). Deep learning for vision-based hand gesture recognition: A review. IEEE Access.
[6] Rautaray, S. S., & Agrawal, A. (2015). Vision-based hand gesture recognition: A review. arXiv. https: //arxiv.org/abs/1507.05765
[7] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing and cloud computing for the Internet of Things: A review. Future Internet, 11(4), 71.
[8] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv. https: //arxiv.org/abs/1905.11946
[9] Warden, P., & Situnayake, D. (2019). TinyML: Machine learning with TensorFlow Lite on Arduino and ultra-low-power microcontrollers. O'Reilly Media.
[10] Xu, D., Li, T., Li, Y., et al. (2020). Edge intelligence: Architectures, challenges, and applications. arXiv. https: //arxiv.org/abs/2003.12172
[11] Gauttam, H., Nain, G., Pattanaik, K. K., & Mendes, P. (2025). Edge-AI: A systematic review on architectures, applications, and challenges. Journal of Network and Computer Applications, 104375.
[12] Ficili, I., Giacobbe, M., Tricomi, G., & Puliafito, A. (2025). From sensors to data intelligence: Leveraging IoT, cloud, and edge computing with AI. Sensors, 25(6), 1763.
[13] Molchanov, P., Gupta, S., Kim, K., & Kautz, J. (2020). A survey on hand gesture recognition datasets and methods. arXiv. https: //arxiv.org/abs/2007.05587