J Inf Process Syst, Vol.21, No.4, pp.371~379, August 2025
https://doi.org/10.3745/JIPS.02.0225 ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

K-Equidistant Partitioning: Enhancing Sensor Data Analysis Using Innovative Data Augmentation Techniques

JeongHyeon Park and Nammee Moon*

Abstract

In this study, *K*-equidistant partitioning (K-EP), a novel data augmentation method, is proposed to address the limitations of sensor data analysis and enhance the performance of behavior classification models. K-EP involves dividing rows of sensor data at equidistant intervals and extracting information from each segment, thereby increasing the size of the dataset by a factor of *K*. This method is based on the sensor minimum warranted frequency hypothesis, which posits that a sampling frequency of 20–40 Hz provides sufficient data for behavioral classification. The effectiveness of K-EP is validated through three experiments, which involve determining the optimal value of *K* for K-EP, comparing K-EP with other data augmentation methods, and assessing the added value of K-EP when combined with other methods. The results indicate that K-EP effectively overcomes the quantitative limitations of sensor data and enhances model robustness. It achieves higher F1-scores than existing methods, indicating that it is an effective data augmentation method for sensor-based behavior classification models.

Keywords

Behavior Classification, CNN, Data Augmentation, Sensor Data

1. Introduction

Recent advancements in sensor technology and machine learning have highlighted the importance of sensor-based behavior analysis in various fields, including healthcare and sports science, where understanding individual behavior patterns is critical. Behavior analysis has applications ranging from daily activity monitoring to complex behavior recognition. However, traditional sensor-based analysis methods face limitations in accuracy and reliability due to the limited quantity and complexity of available data. Therefore, novel approaches to improve the quantity and quality of data are required. This study proposes *K*-equidistant partitioning (K-EP), a novel data augmentation method, to address the quantitative limitations of sensor data and enhance model robustness. This method is based on key findings from sensor-based activity classification research. Previous studies have demonstrated that a sampling frequency between 20 and 50 Hz provides sufficient data for behavior classification [1-6]. Therefore, the sensor minimum warranted frequency (SMWF) hypothesis is proposed, which posits that when the sampling frequency exceeds a certain optimal value, the collected behavioral features are

Manuscript received March 5, 2024; first revision June 21, 2024; accepted October 24, 2024.

* **Corresponding Author:** Nammee Moon (nammee.moon@gmail.com)
Department of Computer Science and Engineering, Hoseo University, Asan, Korea (jh.park970609@gmail.com, nammee.moon@gmail.com)

sufficiently represented, and a further increase in frequency does not considerably affect the performance of the behavior classification model. K-EP involves dividing sensor data rows at equidistant intervals and extracting information from each row, thereby increasing the dataset size by a factor of *K*. This method aims to overcome the limitations of existing data analysis methods and enable accurate behavior recognition.

2. *K*-Equidistant Partitioning

K-EP is a data augmentation method designed to preserve the temporal and frequency characteristics of the original data to address quantitative limitations. Traditional data augmentation methods involve artificially adding, transforming, and generating data similar to the training data or approximating the data distribution using techniques such as generative adversarial networks and variational autoencoders. However, these methods can potentially alter or lose the characteristics of the data. K-EP processes data while preserving the data characteristics, thereby minimizing the loss of data characteristics and maximizing the retention of information from the original data.

The core principle of K-EP is to partition sensor data into *K* equidistant intervals and extract information from each segment, thereby increasing the dataset size by a factor of *K*. In this process, the sampling frequency is based on the minimum frequency changes detectable by the artificial intelligence model. For example, just as the human eye cannot detect rapid movements beyond approximately 60 Hz, the proposed method is based on the concept that additional information beyond a certain frequency does not substantially contribute to behavior classification. The proposed method is based on the SMWF hypothesis, as described in Section 1.

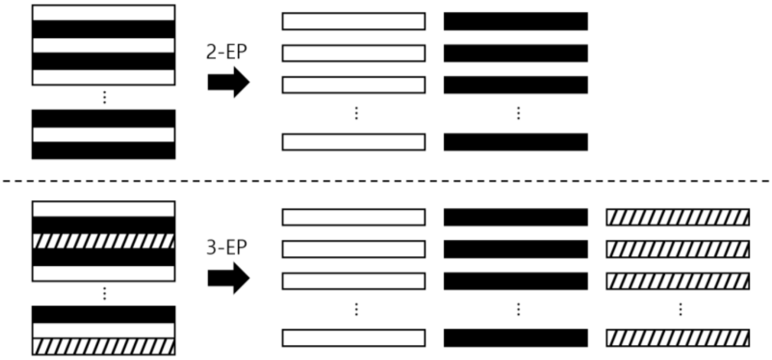


Fig. 1. Example of data partitioning using K-EP with $K = 2$ and $K = 3$.

K-EP involves three stages: *K* selection, partitioning, and synchronization. First, the value of *K* is set to minimize the loss of sensor data characteristics during partitioning. Specifically, *K* is determined considering the SMWF to divide the dataset while minimizing changes in characteristics. In this study, the SMWF was set to 20 Hz based on previous behavior classification research demonstrating that classification model performance decreases at frequencies below 20 Hz. Second, in the partitioning stage, part of the original data is allocated to each equidistant interval, thereby artificially expanding the data while ensuring that the augmented dataset is generated within the range satisfying the SMWF and preserves key data characteristics (Fig. 1). Finally, the synchronization stage consists of merging and training using the original data. This involves adding K empty rows between data segments to which K-EP has been applied, followed by interpolation [7], to ensure that the divided data segments have the same size as the original data. Synchronization is a critical step for ensuring data consistency.

3. Dataset Configuration and Preprocessing Methods

3.1 Dataset

This study employed a dog behavior dataset from Kaggle, containing data collected at a sampling frequency of 100 Hz from 45 medium-to-large dogs equipped with accelerometers and gyroscopes mounted on their necks and backs [8]. The dataset was categorized into six postures and 17 behaviors. For experimental efficiency, these postures and behaviors were grouped into six primary categories: standing, sitting, lying down, walking, playing, and sniffing. Table 1 provides detailed definitions of each category.

Table 1. Behavior definition

Class	Task	Behavior	Definition
Stand	Stand	Standing	The dog is standing with all four paws on the ground, with its torso not touching the ground.
Sit	Sit	Sitting	The dog is sitting with all four paws and its hind quarters on the ground.
Lie down	Lie down	Lying on chest	The dog is lying down with its torso on the ground, with its hips and shoulders at the same height.
Walk	Walk and trot	Walking, pacing, and trotting	The dog is walking, strolling leisurely, or trotting.
Play	Play	Carrying an object, tugging, playing, jumping, and galloping	The dog is engaged in activities including holding an object in its mouth, playing tug-of-war, running, and shaking.
Sniff	-	Eating, drinking, and sniffing	The dog lowers its head below the level of its back and moves its nose close to the ground to sniff. The dog is either standing or moving slowly, with its chest and hindquarters not touching the ground, and may be drinking or picking up and eating food from the ground.

3.2 Data Preprocessing

Data preprocessing was performed separately for each participant to account for variations in maximum movement due to size and weight differences.

Outlier removal: Sensor data may contain outliers due to sensor errors or external factors. These outliers can distort the general pattern of the data, and including them in model training can negatively affect performance. Therefore, the interquartile range (IQR) was employed to detect and remove outliers. The IQR, defined as the difference between the first quartile (Q1) and the third quartile (Q3), represents the middle range of the data (Fig. 2). Values less than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 were considered outliers and removed.

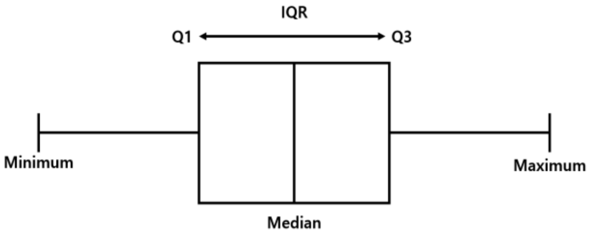
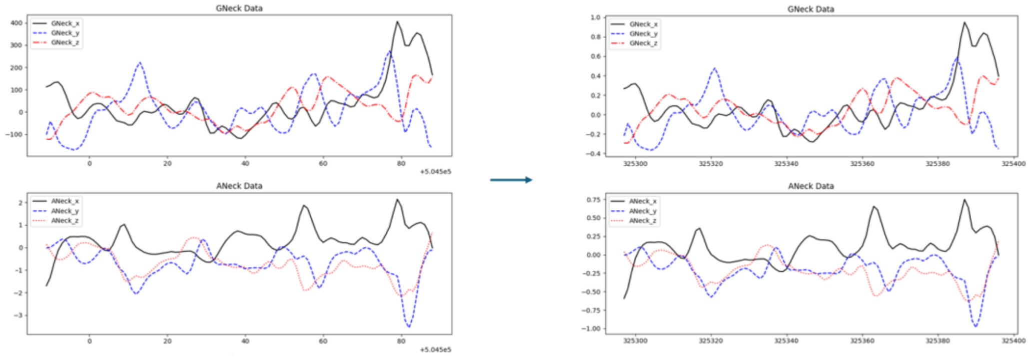


Fig. 2. Interquartile range (IQR).

Data normalization: Data normalization was performed to adjust all features to the same scale, preventing any single feature from disproportionately influencing model training. This step is particularly important when handling data with different units or ranges. In addition, data normalization can increase the convergence speed of machine learning algorithms and reduce local minima problems. In this study, MaxAbsScaler was employed for data normalization, dividing the data by the absolute maximum value of each feature and adjusting all values within the range of −1 to 1. This method is particularly effective when the data distribution is not distorted by extreme values that are far from the median. MaxAbsScaler preserves the data distribution while scaling all features to a uniform range, thereby facilitating efficient training of the deep learning model.

Data sequence generation: Data sequence generation is a critical preprocessing step for time-series data. In this study, the sensor data consisted of continuous multisensor measurements over time, which needed to be converted into a format suitable for processing by deep learning models. In the data sequence generation process, each sensor measurement was segmented based on a predefined time interval, called a window. This window determines the amount of data received by the model at once. For example, a window length of 200 signifies that the model processes 200 consecutive data points per sequence. In this study, the window length was set to 2 seconds, resulting in data sequences with a length of 200 (Fig. 3).



(a)

(b)

Fig. 3. (a) Before and (b) after data processing.

4. Experiment and Results

4.1 Training Dataset Construction

Table 2 presents the training data employed in this experiment. For experimental and evaluation purposes, the dataset was split into training, validation, and test sets in a 6:3:1 ratio.

Table 2. Dataset count

Class	Train	Val	Test	Proportion (%)
Stand	1,913	959	360	9.6
Sit	1,921	959	321	9.5
Lie down	3,350	1,688	526	16.6
Walk	6,555	3,225	1,803	32.4
Play	3,218	1,628	530	16.0
Sniff	3,173	1,606	536	15.8

4.2 Experimental Environment and Methods

Performance evaluation method: The F1-score, which is the harmonic mean of precision and recall, provides a comprehensive assessment of the effectiveness of a model by considering both the number of true positives and the total number of predicted positives. It is particularly useful for binary and multiclass classification, serving as a robust indicator of overall model performance:

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{1}$$

Experimental procedure: All experiments were conducted under identical conditions using the ResNet-18 model trained for 30 epochs.

1) Experiment 1: Determining the optimal *K*

This experiment evaluated the effect of different *K* values on the performance of K-EP. Values of *K* = 2, 3, 4, 5, and 10 were tested, corresponding to sampling frequencies of 50, 33, 25, 20, and 10 Hz, respectively. Examples of data for each experiment are presented in Fig. 4.

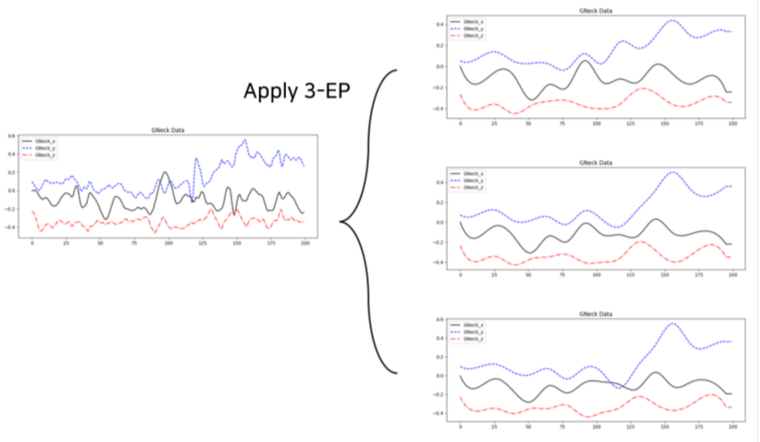


Fig. 4. Example of K-EP with *K* = 3.

2) Experiment 2: Performance comparison with other augmentation methods

In Experiment 2, the performance of K-EP was compared with that of other common augmentation methods for sensor data, such as reversing, inverting, jittering, rotation, time warping, and permutation. This comparison evaluated the effectiveness of K-EP from multiple perspectives, thereby strengthening the results of this study.

3) Experiment 3: Conclusions

In Experiment 3, the performance of models trained on datasets with and without K-EP augmentation was compared and analyzed, using the same augmentation methods as in Experiment 2. This experiment aimed to assess the impact of K-EP augmentation on the generalization ability of the model. In K-EP augmentation, *K* was set to 5 as it yielded the best performance in Experiment 1.

4.3 Experimental Results

4.3.1 Experiment 1

The optimal value of *K* for K-EP was determined using two methods: using only K-EP-augmented data and combining K-EP-augmented data with the original data. Table 3 presents the results.

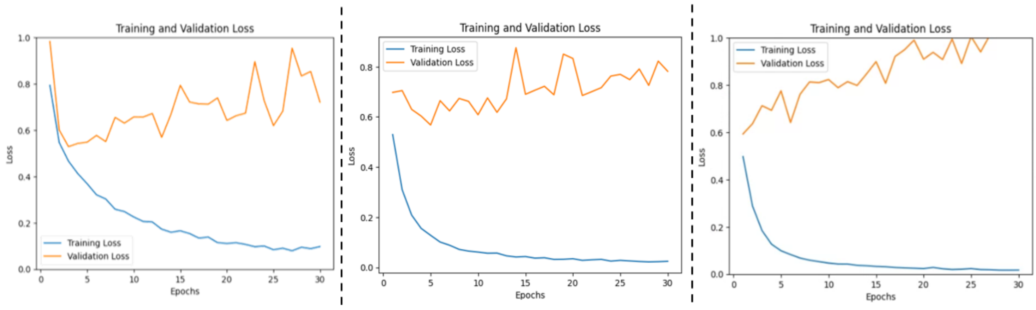
Table 3. Results of Experiment 1

	Original	2-EP	2-EP + Original	3-EP	3-EP + Original	4-EP	4-EP + Original	5-EP	5-EP + Original	10-EP	10-EP + Original
F1-score	0.8247	0.8607	0.8648	0.8680	**0.8803**	0.8494	0.8692	**0.8756**	0.8678	0.8569	0.8734

Best results are highlighted in bold (Kep-only and original+Kep models).

Key observations from Experiment 1 are as follows:

1) Datasets augmented with K-EP consistently achieved higher F1-scores than the original datasets, highlighting the role of K-EP in improving model performance.
2) The SMWF hypothesis was experimentally validated, demonstrating that data quality was preserved when the sampling frequency met the minimum threshold.
3) K-EP-augmented datasets resulted in improved model performance when combined with the original datasets. For example, combining the original dataset with the dataset augmented by K-EP with *K* = 3 considerably increased the F1-score from 0.8680 to 0.8803.
4) The dataset augmented by K-EP with K = 10 achieved a higher F1-score than the original dataset. However, the sampling frequency was reduced to 10 Hz, which did not meet the SMWF threshold. The loss graph in Fig. 5 displays a sharp increase in validation loss, indicating insufficient generalization ability of the model. Therefore, when K-EP is applied with a sampling frequency below the SMWF threshold, essential characteristics of the original sensor data may be lost.



(a) (b) (c)

Fig. 5. Training and validation loss graph: (a) origin, (b) 5-EP, and (c) 10-EP.

4.3.2 Experiment 2

Most augmentation methods led to minimal improvements in model performance, with the exception of the windowing and permutation methods, which led to considerable improvements. However, compared with the augmentation methods frequently employed in artificial intelligence research utilizing sensor data, K-EP exhibited superior performance (Table 4).

Table 4. Results of Experiment 2

	Original	Reverse	Invert	Jitter	VAE	Rotate	Time warp	Permute	Windowing	3-EP
F1-score	0.8247	0.8439	0.8576	0.84	0.8439	0.8317	0.8411	**0.871**	**0.8991**	**0.8803**

Top three results are highlighted in bold (among different augmentation methods).

4.3.3 Experiment 3

The results of Experiment 3 indicated that the model trained on the dataset augmented using methods other than K-EP achieved an F1-score of 0.9307. In contrast, the model trained on the dataset augmented using K-EP achieved an F1-score of 0.9541, demonstrating considerable improvement in model performance. These results suggest that K-EP played a crucial role in enhancing the predictive performance of the model. Notably, the model trained on the K-EP-augmented dataset exhibited lower validation loss (Fig. 6), a key indicator of model robustness. A lower validation loss indicates a better ability of the model to generalize to test data, a reduced risk of overfitting, and more stable overall performance. These results indicate that K-EP can improve the generalization ability of the model.

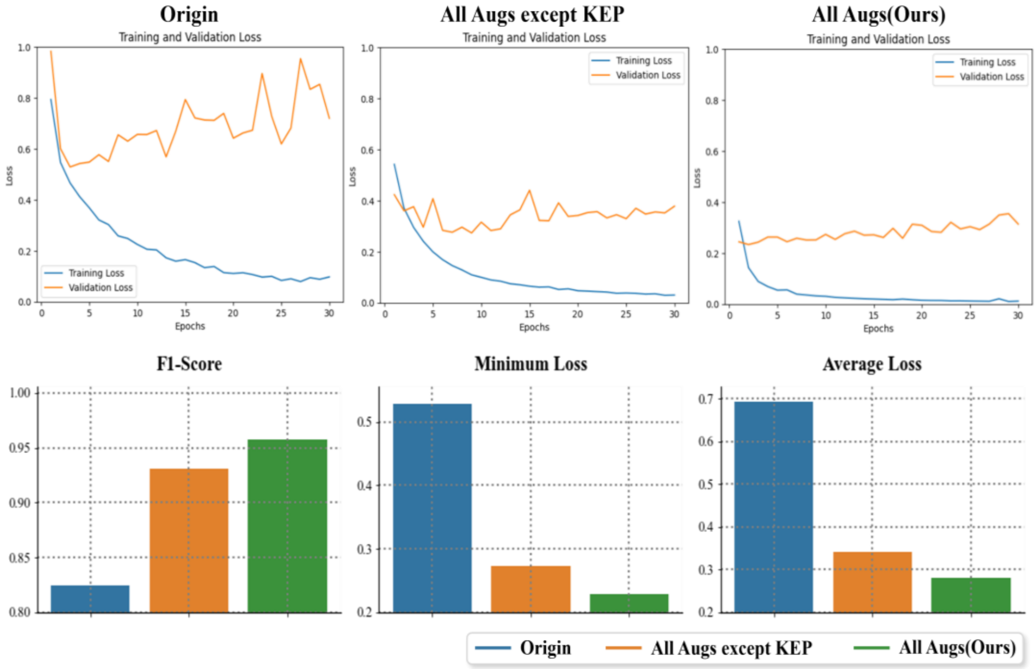


Fig. 6. Training and validation loss (upper) and performance graphs (lower).

5. Conclusion

Datasets augmented by K-EP consistently achieved higher F1-scores than the original datasets, demonstrating the effectiveness of K-EP in improving model performance. In addition, the SMWF hypothesis was experimentally validated, confirming that K-EP preserved essential sensor data characteristics while increasing the dataset size. Furthermore, K-EP enhanced the model's generalization ability, indicating its potential to greatly improve performance and providing a solid foundation for future research.

Because K-EP augmentation was applied by a factor of K to all classes, the amount of data for the Walk class was, on average, twice as much as for other classes, potentially introducing data bias that may have negatively affected the model. To address this problem, additional experiments will be conducted in future studies.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A2C2011966).

References

[1] J. A. Santoyo-Ramon, E. Casilari, and J. M. Cano-Garcia, "A study of the influence of the sensor sampling frequency on the performance of wearable fall detectors," *Measurement*, vol. 193, article no. 110945, 2022. https://doi.org/10.1016/j.measurement.2022.110945

[2] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," in *Proceedings of International Workshop on Wearable and Implantable Body Sensor Networks (BSN)*, Cambridge, MA, USA, 2006, pp. 113-116. https://doi.org/10.1109/BSN.2006.6

[3] S. D. Bersch, D. Azzi, R. Khusainov, I. E. Achumba, and J. Ries, "Sensor data acquisition and processing parameters for human activity classification," *Sensors*, vol. 14, no. 3, pp. 4239-4270, 2014. https://doi.org/10.3390/s140304239

[4] Z. Pan, H. Chen, W. Zhong, A. Wang, and C. Zheng, "A CNN-based animal behavior recognition algorithm for wearable devices," *IEEE Sensors Journal*, vol. 23, no. 5, pp. 5156-5164, 2023. https://doi.org/10.1109/JSEN.2023.3239015

[5] A. Eerdekens, A. Callaert, M. Deruyck, L. Martens, and W. Joseph, "Dog's behaviour classification based on wearable sensor accelerometer data," in *Proceedings of 2022 5th Conference on Cloud and Internet of Things (CIoT)*, Marrakech, Morocco, 2022, pp. 226-231. https://doi.org/10.1109/CIoT53061.2022.9766553

[6] J. Kim, H. Kim, C. Park, and N. Moon, "Deep learning-based pet monitoring system and activity recognition device," *Journal of the Korea Society of Computer and Information*, vol. 27, no. 2, pp. 25-32, 2022. https://doi.org/10.9708/jksci.2022.27.02.025

[7] K. Kodera, A. Nishitani, and Y. Okihara, "Cubic spline interpolation based estimation of all story seismic responses with acceleration measurement at a limited number of floors," *Journal of Structural and Construction Engineering*, vol. 83, no. 746, pp. 527-535, 2018. https://doi.org/10.3130/aijs.83.527

[8] A. Vehkaoja, S. Somppi, H. Tornqvist, A. V. Cardo, P. Kumpulainen, H. Vaataja, et al., "Description of movement sensor dataset for dog behavior classification," *Data in Brief*, vol. 40, article no. 107822, 2022. https://doi.org/10.1016/j.dib.2022.107822

JeongHyeon Park **https://orcid.org/0000-0002-3832-4946**

He received his B.S. degree from the Department of Computer Science and Engineering, Hoseo University, in 2023. Since March 2023, he has been with the Department of Computer Science and Engineering, Hoseo University, as a M.S. candidate. His current research interests include big data processing and analysis.

Nammee Moon **https://orcid.org/0000-0003-2229-4217**

She received her B.S., M.S., and Ph.D. degrees from the School of Computer Science and Engineering, Ewha Womans University, in 1985, 1987, and 1998, respectively. She served as an assistant professor at Ewha Womans University from 1999 to 2003 and then as a professor of digital media, Graduate School of Seoul Venture Information, from 2003 to 2008. Since 2008, she has been a professor of computer information at Hoseo University. Her current research interests include social learning, human–computer interaction, user-centric data, and big data processing and analysis.