JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 41, XXX-XXX (2025)
DOI: 10.6688/JISE.202511_41(6).0001

Enhanced Pet Activity Recognition With ViT and Diffusion Model through Sensor to Image Encoding

JEONG-HYEON PARK AND NAMMEE MOON+
Department of Computer Science and Engineering
Hoseo University
Asan, 31498 Korea
E-mail: jh.park970609@gmail.com; nammee.moon@gmail.com

Time-series sensor data collected from wearable devices has led to significant advancements in health monitoring through the application of AI. However, the complexity and noise sensitivity of this data have slowed progress, leading to recent studies exploring the transformation of data into image formats to apply image processing techniques. In this study, we aim to classify pet activities by transforming sensor data into various image formats, rather than in the traditional time-series format, and applying data augmentation using diffusion models followed by the application of Vision Transformer (ViT). The datasets used include a self-collected dataset and the Dog Behavior Analysis (DBA) dataset from Kaggle, unified into five activity classes. The sensor data underwent preprocessing steps such as outlier removal, missing value interpolation, and normalization before being segmented into 2.56-second sequences to form sequence data. These sequence data were then transformed into images to be used as inputs for the classification model. Experimental results showed that when comparing the performance of the original sensor data as input to a transformer model with the proposed Time Series Image Rescaling (TIR) format transformed data as input to the ViT model, the TIR format achieved an F1 score improvement of 0.0901 on our dataset and 0.0726 on the DBA dataset. Additionally, the ViT model outperformed various CNN models, achieving the highest performance on both datasets. To address data bias, sensor images were augmented using the diffusion model. The augmented data were added to the existing dataset and used as input for the ViT model, resulting in an F1 score improvement of 0.085 on our dataset and 0.0618 on the DBA dataset. The results of this study demonstrate that transforming time-series sensor data into sensor images and applying advanced image processing techniques can enhance Pet Activity Recognition (PAR) performance. Future research will aim to collect data from various pets, such as cats, in addition to dogs, to recognize and predict the behavior of a broader range of species.

***Keywords*:** diffusion model, sensor to image encoding, vision transformer, behavior classification

1. INTRODUCTION

Sensor data collected through wearable devices has driven advancements in health monitoring and healthcare by integrating with artificial intelligence technologies [1-3]. While wearable devices provide a variety of data for analysis and application, most sensor data is offered in a time-series format. Time-series data, which handles continuous data that changes over time, is inherently complex and sensitive to noise. Consequently, advancements in time-series data analysis have lagged behind those in image processing, limiting its full potential. Recent studies have explored various methods to utilize image

Received March 25, 2024; revised June 14, 2024; accepted August 8, 2024.
Communicated by Ji Su Park.
+ Corresponding author.

processing techniques by converting sensor data into image formats [4-6].

In this study, we aim to classify pet activities by converting sensor data from wearable devices into image formats and applying the latest image processing technologies, specifically diffusion models and Vision Transformer (ViT) [7-10]. The conversion of sensor data into image formats is achieved using both existing techniques and the simple conversion methods proposed in this study [4-6]. The transformed sensor images are then classified using ViT, and their performance is evaluated using various metrics.

In this study, we aim to improve the performance of behavior classification models by combining augmented sensor images with the existing training dataset. Traditional sensor data analysis approaches are beneficial but have efficiency limitations. By integrating the improvements offered by optimal sensor-to-image encoding methods and diffusion models, we expect a significant enhancement in the accuracy and reliability of these classification models. These advancements can have a broad impact on more accurate health monitoring, early detection of health anomalies, and personalized medical solutions. Additionally, it is expected to improve the performance of various sensor data classification problems.

2. RELATED WORK

2.1 Transforming Sensor Data to Image Format

Recent studies have been conducted on transforming time-series data into image format for use with deep learning algorithms like CNN (Convolutional Neural Networks) [11-15]. This approach effectively contributes to extracting complex patterns that are difficult to capture with traditional time-series analysis methods. According to a study analyzing vibration measurement data collected from Airbus SAS's helicopter flight tests, six methods of transforming time-series data into images (Gramian Angular Field, Markov Transition Field, Recurrence Plot, Gray Scale Encoding, Spectrogram, and Scalogram) have proven effective in detecting anomalies [16]. Particularly, these methods have shown significant progress in robustness against variability in large datasets.

Additionally, a study titled 'Towards Improved Human Action Recognition' proposes a novel method of transforming depth data into sequential front images and inertial data into signal images [17]. This method enhanced the accuracy of human action recognition by fusing depth and inertial sensor data and demonstrated superior performance in experiments integrating CNN, SVM, and Softmax classifiers.

These studies validate that transforming time-series data into images can improve model performance compared to traditional methods and open up possibilities for deeper insights and more powerful predictions in data analysis by applying image processing algorithms like CNN.

2.2 Diffusion Model

Originating in the field of statistical physics, diffusion models have been extensively applied in various domains, notably in image processing and generation [7]. These models effectively facilitate data augmentation and feature extraction by progressively adding

noise to the original data and subsequently reversing the process through a simulation that gradually removes noise, restoring the data to its original form. In recent years, diffusion models have distinguished themselves within the image processing community due to their exceptional capability in generating high-quality, diverse images and their potential in various other applications. One of the most notable applications of diffusion models is in the realm of image generation and enhancement. A recent study has demonstrated their superior performance compared to existing generative model algorithms, such as GANs [18]. Furthermore, research involving the generation of synthetic sensor data for pet Behavior Prediction using GAN algorithms has shown improvements in behavior classification models [19].

In this study, we aim to employ diffusion models to generate synthetic samples of sensor data. This approach is anticipated to advance the current state of sensor data analysis by leveraging the robust capabilities of diffusion models in generating realistic and diverse datasets, thus enhancing the performance of classification models and contributing to a broader understanding of sensor-based monitoring systems.

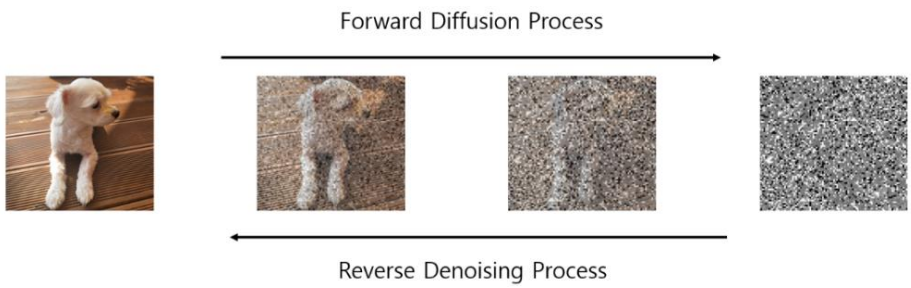

Fig. 1. Example diffusion model process.

2.3 Vision Transformer (ViT)

Vision Transformer (ViT) is an innovative deep learning model that has recently gained attention in the field of computer vision for its exceptional performance in image processing [10]. Unlike traditional CNN-based models, ViT utilizes a transformer architecture to handle image data. This model, originally successful in the natural language processing (NLP) domain, has been adapted for image processing, demonstrating superior performance in various tasks such as image classification. A key feature of ViT is its method of processing images by dividing them into patches, performing linear transformations on each patch, and inputting them into a transformer encoder for feature extraction that considers the global context. This approach contrasts with CNN models that primarily focus on local features, as ViT allows for a comprehensive analysis of the entire image. Recent studies have shown that transforming sensor data into image formats and using them as training data for ViT can enhance the accuracy of Human Activity Recognition (HAR). ViT has demonstrated its ability to recognize complex patterns and maintain high performance across large datasets with diverse variations, outperforming traditional CNN-based approaches.

In this study, we aim to improve the accuracy of predicting Pet Activity Recognition (PAR) using 3-axis sensor data from pets by employing ViT.

3. PAR WITH VIT AND DIFFUSION MODEL THROUGH SENSOR TO IMAGE ENCODING

This study proposes a method for PAR using ViT through data augmentation with diffusion models and time series data encoding. The overall process is illustrated in Fig. 2.

First, the preprocessing of pet behavior sensor data is carried out, followed by encoding the data into an image format. Once the data is transformed into images, a diffusion model is trained using this data. The trained diffusion model is then used to adjust the biased training data to ensure uniformity. Finally, ViT is employed for PAR.

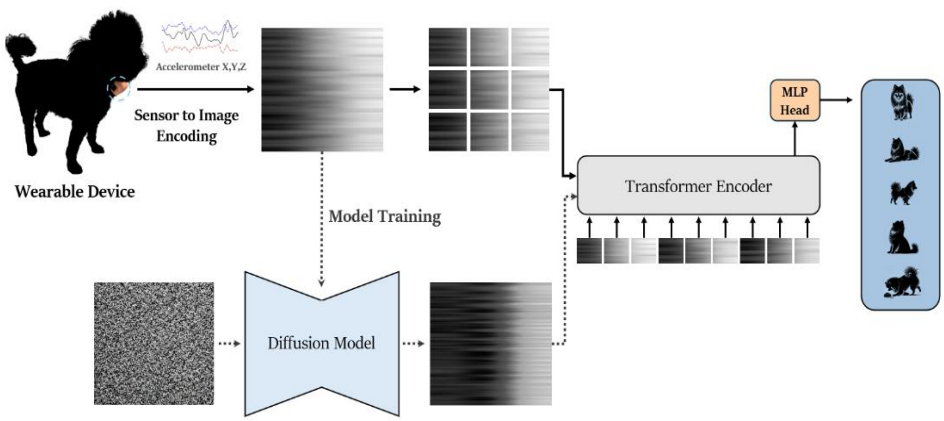

Fig. 2. PAR with ViT and diffusion models through sensor to image encoding process.

3.1 Dataset

In this study, we utilized the dog behavior dataset from Kaggle and our own collected dog behavior dataset [20, 21]. Due to the different collection environments and devices used for the two datasets, we did not merge the data. However, to evaluate the performance of the proposed methodologies in the paper, we set up the experimental environments to be as similar as possible.

(A) Our Dataset

The dataset was collected using a wearable device created with a 9-axis sensor-based Printed Circuit Board (PCB). The data collected by the wearable device is transmitted to an ODROID, where the sensor data is periodically sent to a web database. Excluding the collar, the device weighs about 28g, and the case was produced using a 3D printer. The data collection targeted ten small to medium-sized dogs, ranging in weight from 1.5 kg to 14.5 kg, and aged from 12 to 150 months, with the device attached to their necks. The sensor data was recorded using a 9-axis sensor (accelerometer, gyroscope, magnetometer) at a frequency of 50Hz. The dataset categorized the dogs' behaviors into nine different posture behaviors. However, to align with a similar environment provided by a Kaggle dataset on dog behavior, it was condensed into five basic actions, each defined in detail in Table 1.

Fig. 3. Our dataset collection process.

(B) Dog Behavior Analysis Dataset (DBA Dataset)

The DBA dataset was collected using two ActiGraph GT9X Link devices, each equipped with a 3-axis accelerometer and a 3-axis gyroscope [20, 21]. The collection targeted 45 mediums to large dogs, aged between 16 and 116 months and weighing between 13 kg and 41 kg. The devices were attached to both the neck and back areas of each dog. The sensor data was recorded at a frequency of 100 Hz using a 6-axis (accelerometer and gyroscope) sensor setup. The dataset categorized the dogs' behaviors into 6 postures and 17 movements. However, to simulate an environment similar to that of the directly collected data, it was condensed into 5 basic actions, each of which is detailed in Table 1.

Table 1. Behavior definition table.

Class	Definition
Stand	The hind legs or all four legs are touching the ground, with at least two legs standing in a straight line.
Sit	The front legs are in contact with the ground in a straight line, while the hind legs are bent. Additionally, the tail or the rear end is touching the ground.
Lie down	All four legs are not touching the ground, or all legs are bent or extended in a non-vertical position relative to the ground. The belly or sides are in contact with the ground.
Walk	The animal is moving forward without all four legs simultaneously touching the ground. This forward movement occurs regardless of speed.
Eat & Sniff	The head is positioned lower than the back, with the nose moving close to the ground. This includes sniffing, chewing, drinking, or any form of ingestion.

3.3 Data Processing

Data preprocessing is performed individually for each subject, considering the differences in maximum movement due to varying sizes and weights of the dogs.

(A) Outlier Removal

Sensor data can have outliers due to various reasons, such as sensor errors or exceptional events. These outliers can distort the general pattern of the data and including them in model training can degrade performance. This study uses the IQR method to detect and remove outliers. IQR represents the difference between the first and third quartiles of the data, defining the data's middle range. Values below 1.5 IQR from Q1 or above 1.5 IQR from Q3 are considered outliers and removed.

(B) Missing Value Interpolation

Missing values in sensor data can occur due to sensor malfunctions, transmission errors, or other disruptions. These missing values can lead to inaccurate analysis and model

training issues. This study employs cubic spline interpolation to estimate and fill in these missing values. Cubic spline interpolation uses piecewise cubic polynomials to estimate the missing values, ensuring a smooth and continuous curve that passes through the known data points. By filling in these gaps with cubic spline interpolation, the integrity of the time series data is maintained, allowing for more accurate and reliable analysis and model training.

(C) Data Normalization

Data normalization adjusts all features to the same scale, preventing any single feature from disproportionately influencing model training. This is particularly important when dealing with data with different units or ranges. Additionally, normalized data can enhance the convergence speed of machine learning algorithms and reduce local minima problems. This study uses the Robust Scaler for data normalization, dividing the data by the absolute maximum value of each feature, adjusting all values within a range of -1 to 1. This method is particularly useful when the data distribution is not distorted by extreme values far from the center. Robust Scaler maintains the distribution of data while ensuring that all features are on the same scale, aiding the deep learning model in more efficient training.

(D) Data Sequence Creation

Creating data sequences is a crucial preprocessing step for time series data. The sensor data used in this study consists of continuous streams of various sensor measurements over time, making it essential to transform this data into a format that can be processed by deep learning models. This transformation process is referred to as data sequence generation. Data sequences are created by dividing continuous sensor data into fixed time intervals, known as windows. For example, setting a window length of 256 means that the model receives 256 consecutive data points.

In this study, a window length of 2.56 seconds is set, with the dataset directly collected for each data point being 128, and the Dog Behavior Analysis Dataset from Kaggle being set to 256.

Table 2. Our dataset count.

No. (Label)	Class	Train	Val	Test	Proportion
0	Walk	763	220	120	7.32%
1	Lie down	3,727	1,068	544	36.01%
2	Sit	2,897	819	395	27.73%
3	Sniff & Eat	1,116	330	157	10.81%
4	Stand	1,875	543	271	18.14%

Table 3. DBA dataset count.

No. (Label)	Class	Train	Val	Test	Proportion
0	Walk	8,800	2,537	1,240	48.78%
1	Lie down	3,037	854	425	16.74%
2	Sit	1,703	497	248	9.49%
3	Sniff & Eat	2,795	801	384	15.44%
4	Stand	1,712	493	256	9.55%

3.4 Sensor Data to Image Encoding

To efficiently apply ViT and diffusion models, which are widely used in the field of image processing, it is essential to encode sensor data into image formats. For ViT in particular, ensuring that the input image size is 224×224 is crucial. This is because ViT processes images by dividing them into fixed-size patches. Maintaining a consistent input size is necessary as it directly impacts the model's performance and training efficiency.

This study explores six methods to encode sensor data into image formats. Specifically, we investigate Time Series Image Rescaling, Outer Product, and average matrix transformation, as proposed in this paper, along with widely used methods such as Gramian Angular Field, Markov Transition Field, and Recurrence Plot. By resizing the images generated by each method to 224×224, we aim to maximize the application of advanced image processing technologies like ViT.

1) Time series Image Rescaling (TIR)

Time series Image Rescaling (TIR) transforms sequenced time series data into a simple rectangular image format, optimizing it for ViT models in an intuitive and quick manner. This method involves setting the target image size (H, W) and using the Image.resize function from the Python Imaging Library (PIL) to adjust the array size. This approach visualizes time series patterns and generates images that can be used as inputs for deep learning models. The main advantage of TIR transformation lies in its simplicity and intuitiveness. It converts time series data into images without complex preprocessing, making it easier to analyze patterns. Additionally, the generated images can be utilized as inputs for other image-based deep learning models, such as Convolutional Neural Networks (CNN), enabling various analysis and prediction tasks.

2) Outer Product (OP)

The Outer Product (OP) method calculates the outer product of a time series to transform one-dimensional time series data into two dimensions. This method generates a matrix where each element is the product of two points in the time series, effectively capturing the interactions between different time points. The resulting matrix can be used as input for machine learning models. This matrix captures the relationships between all pairs of points in the time series, providing a comprehensive representation of the data. The outer product is defined by Eq. (1) when the time series is $x = [x_1, x_2, \ldots, x_n]$

$$X = x \otimes x = \begin{bmatrix} x_1 \cdot x_1 & x_1 \cdot x_2 & \cdots & x_1 \cdot x_n \\ x_2 \cdot x_1 & x_2 \cdot x_2 & \cdots & x_2 \cdot x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_n \cdot x_1 & x_n \cdot x_2 & \cdots & x_n \cdot x_n \end{bmatrix}. \tag{1}$$

3) Average Matrix Transformation (AMT)

The Average Matrix Transformation (AMT) method is used to convert one-dimensional time series data into two dimensions by calculating the average of each pair of elements in the time series data. This method computes the average values between two points in the time series, thereby generating a new matrix that effectively captures the interactions

within the time series data. The resulting matrix can be used as input for machine learning models. This matrix captures the relationships between all pairs of points in the time series, providing a comprehensive representation of the data. The AMT is defined as in Eq. (2).

$$A_{ij} = \frac{x_i + x_j}{2} \tag{2}$$

(4) Gramian Angular Field (GAF)

The Gramian Angular Field (GAF) transformation converts time series data into images by first transforming the data into polar coordinates and then calculating the cosine of the angles between the data points to create a Gramian matrix [22]. This matrix preserves temporal dependencies, allowing the visualization of time series patterns as images. The GAF transformation proceeds through normalization, conversion to polar coordinates, and generation of the Gramian matrix. Normalization scales the time series data values to a range between −1 and 1. The polar conversion transforms these normalized values into angles, where each angle corresponds to a data point in the time series. The Gramian matrix is then generated by calculating the cosine values between these angles. Each element of the matrix represents the correlation between the data values at two different time points.

In this study, since the data normalization has already been performed during the preprocessing stage, the GAF transformation involves only the polar conversion and Gramian matrix generation steps.

$$\phi_i = \arccos(S_i) \tag{3}$$

$$G_{ij} = \cos(\phi_i + \phi_j) \tag{4}$$

(5) Markov Transition Field (MTF)

The Markov Transition Field (MTF) is a technique designed to capture the changing patterns within time series data by reconstructing the data into a Markov matrix using the transition probabilities between states at each point in time. In this process, each state is assigned a number from 1 to n, and each element ω_{ij} of the Markov matrix represents the probability of transitioning from state i to state j. These transition probabilities must be normalized so that their sum equals 1, reflecting the overall dynamic structure of the data. The MTF methodology is particularly useful for analyzing or comparing latent patterns within time series data by providing a visualization of the temporal structure in the form of an image.

$$M = \begin{pmatrix} \omega_{ij} \mid x_1 \in q_i, x_1 \in q_j \cdots \omega_{ij} \mid x_1 \in q_i, x_n \in q_j \\ \vdots \qquad \ddots \qquad \vdots \\ \omega_{ij} \mid x_n \in q_i, x_1 \in q_j \cdots \omega_{ij} \mid x_n \in q_i, x_n \in q_j \end{pmatrix} \tag{5}$$

(6) Recurrence Plot (RP)

The Recurrence Plot (RP) is another innovative method used to analyze time series data [23]. This approach visualizes the recurrences of states within a time series, offering insights into the dynamic behavior of the system being studied. The key idea of RP is to plot a two-dimensional square matrix, where each axis represents the entire time series. Points within the matrix are marked if the state at a specific time is similar to the state at

another time, based on a predefined criterion of similarity. Recurrences in the data are visualized as patterns of dots in the plot, which can reveal important information such as cyclic behavior, periodicity, or sudden changes in the time series. The recurrence plot is mathematically defined by the Eq. (6).

$$R_{ij} = \Theta(\epsilon - ||X_i - X_j||) \tag{6}$$

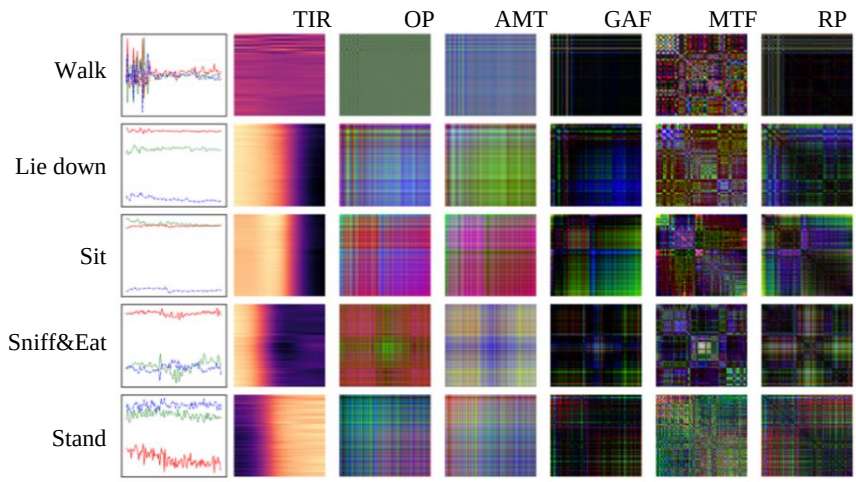

Fig. 4. Image encoding example for each class.

3.5 Experiment Procedure

To evaluate the performance of the model, precision, recall, and F1 score are used. Precision measures the accuracy of positive predictions, while recall assesses the model's ability to identify all relevant instances. The F1 score, as a harmonic mean of precision and recall, provides a balanced evaluation of the model's effectiveness. These metrics together offer a comprehensive assessment of the model's performance for both binary and multi-class classification.

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \tag{7}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

(A) Performance experiments by data format

In this experiment, we assess the performance of each image encoding method to validate the effectiveness of sensor data formats for PAR. We use a total of seven formats: Original, TIR, OP, AMT, GAF, MTF, and RP. Each format, except for the original data, is trained using the ViT model. The original data, being in time series format rather than image format, is trained using a standard Transformer model instead of ViT. The performance of each format is evaluated based on accuracy, precision, recall, and F1 scores.

(B) Performance experiments for each image classification model

In this experiment on image classification model performance, we aim to conduct a comparative analysis using various models. This process involves transforming data into the optimal form selected from previous experiments. We will utilize deep learning models such as ViT, ResNet, EfficientNet, VGG, DenseNet, and Inception to carry out PAR. This stage is crucial for assessing the performance of the proposed ViT in PAR. The performance of all models will be evaluated based on precision, recall, and F1 scores.

(C) Performance measurement based on augmentation using diffusion model

In the final experiment, the data is transformed using the TIR format, which showed the best performance in previous experiments, and this transformed data is used to train the diffusion model. The diffusion model focuses on learning the distribution of the data and generating new images that resemble the actual data. Using the trained diffusion model, new sensor images are generated for underrepresented classes in the existing dataset. This process aims to address the data imbalance problem and help the model learn from a more diverse set of data. The generated augmented images are added to the existing training dataset, forming a new augmented training dataset. This augmented dataset is then used to perform PAR with the ViT model. The performance of the model trained with the augmented dataset is evaluated based on accuracy, precision, recall, and F1 score. Additionally, a comparative analysis is conducted to examine the effect of image augmentation on classification accuracy. This comparative analysis plays a crucial role in assessing whether data augmentation indeed enhances the model's generalization capabilities and provides better predictive performance in various scenarios.

4. EXPERIMENT

This study was implemented using the Python library Pytorch. Table 4 lists the detailed specifications of the experimental environment.

Table 4. Experiment environment.

CPU	Intel I9-9900k
GPU	GeForce Titan RTX 2way
RAM	64G
CUDA	11.7
cuDNN	8.4.1.50
Python	3.9.6
Pytorch	1.13.0

4.1 Performance Experiments by Data Format

In this experiment, the optimizer for PAR set AdamW to a learning rate of 0.00005, and CrossEntropyLoss was applied as the loss function. To prevent overfitting, L2 regulation and label smoothing were applied. Performance indicators were measured based on precision, recall and F1 score.

Table 5. PAR performance by data format.

Dataset	Format	Model	Precision	Recall	F1-Score
Our Dataset	Original	Transformer	0.80	0.80	0.7943
	TIR	**ViT/B_16**	**0.88**	**0.88**	**0.8844**
	OP	ViT/B_16	0.83	0.83	0.8310
	AMT	**ViT/B_16**	**0.88**	**0.88**	**0.8804**
	GAF	ViT/B_16	0.69	0.69	0.6903
	MTF	ViT/B_16	0.61	0.62	0.6113
	RP	ViT/B_16	0.75	0.75	0.7462
DBA Dataset	Original	Transformer	0.84	0.84	0.8384
	TIR	**ViT/B_16**	**0.91**	**0.91**	**0.9110**
	OP	**ViT/B_16**	**0.91**	**0.91**	**0.9124**
	AMT	ViT/B_16	0.89	0.88	0.8838
	GAF	ViT/B_16	0.81	0.82	0.8144
	MTF	ViT/B_16	0.80	0.81	0.8098
	RP	ViT/B_16	0.80	0.81	0.8028

Performance experiments by data format revealed that applying the TIR and AMT methods to ViT resulted in the best performance across both datasets. Two key insights can be drawn from this observation.

First, it appears that minimal transformation of the original data yields the best performance. This suggests that methods such as GAF, MTF, and RP, which are commonly used for radio and voice data operating at kHz frequencies, may not be suitable for wearable device sensor data. Wearable sensor data typically operates at lower sampling rates, such as 25-100Hz, and converting these small sample rate time series data into images using traditional methods may not be effective.

Second, the performance of PAR improves when the original data is transformed into TIR, OP, or AMT formats and applied to ViT, compared to training a Transformer model on the raw data. Transformer models were originally designed for natural language processing (NLP), focusing on understanding long sequences and context, with the primary goal of predicting the next word in a sequence. These models are adept at capturing dependencies and relationships within sequential data.

However, for behavior prediction using sensor data, the data is usually sequenced, which means it does not typically involve long sequences. Additionally, when sensor data is converted into TIR, OP, or AMT formats, it is effectively transformed into a structured format that ViT can process more efficiently. ViT is proficient at learning complex patterns overall, which is likely why performance is higher with this structured format.

4.2 Performance Experiments for Each Image Classification Model

Performance experiments were conducted using various image classification models on two datasets: Our Dataset and the DBA Dataset. Each model utilized pre-trained models with the TIR format. The following is a detailed analysis of the results obtained. The ViT/B_16 model consistently delivered the highest performance across both datasets. In Our Dataset, ViT/B_16 outperformed other models by a margin of up to 0.077 in F1-score, and at a minimum, the difference was 0.0148. For the DBA Dataset, the ViT/B_16 model showed a maximum performance difference of 0.0155 and a minimum of 0.0013 compared

to other CNN-based models. Performance indicators for each CNN model are detailed in Table 6, illustrating the comparative effectiveness and reliability of these models in handling image classification tasks.

Table 6. PAR performance by classification model.

Dataset	Model	Precision	Recall	F1-Score
Our Dataset	**ViT/B_16**	**0.88**	**0.88**	**0.8844**
	ResNet-18	0.84	0.84	0.8353
	ResNet-50	0.84	0.84	0.8417
	EfficientNet,	0.80	0.81	0.8104
	VGG11	0.87	0.87	0.8696
	DenseNet	0.86	0.86	0.8593
	Inception	0.80	0.81	0.8074
DBA Dataset	**ViT/B_16**	**0.91**	**0.91**	**0.9110**
	ResNet 18	0.92	0.90	0.9092
	ResNet 50	0.91	0.91	0.9094
	EfficientNet b0	0.89	0.90	0.8994
	VGG11	0.89	0.89	0.8955
	DenseNet 121	0.90	0.91	0.9097
	Inception v3	0.91	0.90	0.9037

4.3 Performance Measurement Based on Augmentation using Diffusion Model

In the initial experiment, sensor images encoded with TIR were effectively learned through a diffusion model. The diffusion model was constructed using the denoising_diffusion_pytorch library in Python. The generated data is as shown in Fig. 5. This augmentation process generated an additional 1,792 to 2,560 samples per label for the classes with fewer data, ensuring that it did not exceed the data of the most abundant class. These were then integrated into the training dataset. The numbers of data per class before and after augmentation in the training dataset is shown in Table 7.

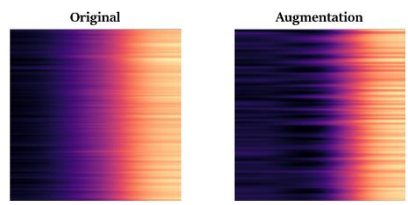

Fig. 5. TIR format data and below a sample of data generated by the diffusion model.

Table 7. Number of data before and after augmentation for each data.

Dataset	Before	After	Dataset	Before	After
Our Dataset	763	3,323	DBA Dataset	8,800	8,800
	3,727	3,727		3,037	5,597
	2,897	2,897		1,703	4,263
	1,116	3,676		2,795	5,355
	1,875	3,667		1,712	4,272

PAR was performed using the ViT architecture, and a comparative evaluation of model performance before and after augmentation was conducted. In our dataset, the F1-Score increased from 0.8844 to 0.9694, showing a performance improvement of 0.085. Similarly, in the DBA Dataset, the F1-Score improved from 0.911 to 0.9728, indicating a significant enhancement of 0.0618. As depicted in Fig. 6. the post augmentation confusion matrices show higher diagonal values, which indicates an increase in correct predictions and a dramatic reduction in misclassifications. This underscores the effectiveness of data augmentation techniques in resolving dataset imbalance issues and enhancing prediction.

Table 8. Classification performance for each behavior.

Dataset	No. (Label)	Behavior	Precision	Recall	F1-Score
Our Dataset	0	Walk	0.93	0.98	0.96
	1	Lie down	0.99	0.97	0.98
	2	Play	0.96	0.98	0.97
	3	Stand	0.98	0.96	0.97
	4	Sit	0.96	0.96	0.96
DBA Dataset	0	Walk	0.99	0.99	0.99
	1	Lie down	0.94	0.97	0.96
	2	Play	0.93	0.92	0.93
	3	Stand	0.99	0.98	0.99
	4	Sit	0.93	0.89	0.91

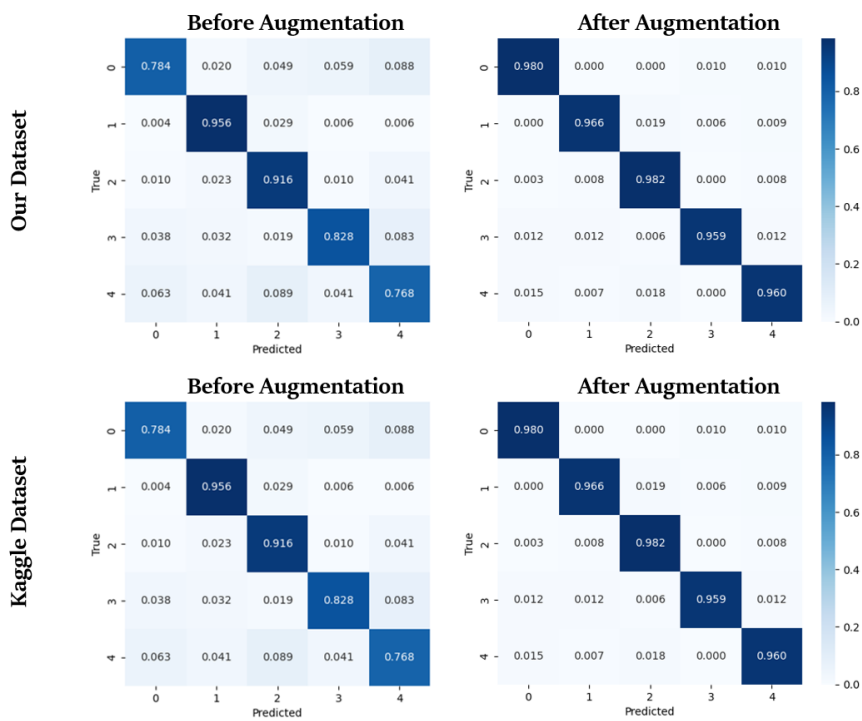

Fig. 6. Confusion matrix graph before and after augmentation.

5. CONCLUSION

This study aims to perform PAR using a 3-axis accelerometer collected from wearable devices for pets. The datasets used include a dataset directly collected and the DBA dataset from Kaggle, both unified into five activity classes. The sensor data underwent outlier processing, missing value interpolation, and normalization before being divided into 2.56-second sequences to form sequence data. These sequence data were then transformed into images to be used as input for the classification model. When the original sensor data were used as input to the transformer model, the performance was compared to using the proposed TIR format transformed data as input to the ViT model. The TIR format showed a performance improvement of 0.0901 in F1-score on our dataset and 0.0726 on the DBA dataset. To address data bias, sensor images were augmented using the diffusion model. The augmented data were added to the existing dataset and used as input for the ViT model to perform PAR. This approach showed a performance improvement of 0.085 in F1-score on our dataset and 0.0618 on the DBA dataset.

The results of this study demonstrate the significance of applying advanced image processing techniques to sensor data image transformation to enhance PAR performance. Future research will involve collecting data from various pets, such as cats, in addition to dogs, to recognize and predict the behavior of a broader range of species.

REFERENCES

1. W. Ding, "Role of sensors based on machine learning health monitoring in athletes' wearable heart rate monitoring," *Human-Centric Computing and Information Sciences*, Vol. 13, 2023, pp. 1-16.
2. J. K. Kim, K. B. Lee, and S. G. Hong, "Cognitive load recognition based on T-test and SHAP from wristband sensors", *Human-centric Computing and Information Sciences*, Vol. 13, 2023, pp. 1-14.
3. D. Alghazzawi, O. Rabie, O. Bamasaq, A. Albeshri, and M. Z. Asghar "Sensor-based human activity recognition in smart homes using depthwise separable convolutions," *Human-centric Computing and Information Sciences*, Vol. 12, 2022, pp. 1-19.
4. C. L. Yang, C. Y. Yang, Z. X. Chen, and N. W. Lo, "Multivariate time series data transformation for convolutional neural network," in *Proceedings of IEEE/SICE International Symposium on System Integration*, 2019, pp. 188-192.
5. S. Barra, S. Carta, A. Corriga, A. S. Podda, and D. R. Recupero "Deep learning and time series-to-image encoding for financial forecasting," *IEEE/CAA Journal of Automatica Sinica*, Vol. 7, 2020, pp. 683-692.
6. Z. Ahmad and N. Khan, "Inertial sensor data to image encoding for human action recognition," *IEEE Sensors Journal*, Vol. 21, 2021, pp. 10978-10988.
7. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, Vol. 33, 2020, pp. 6840-6851.
8. S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet "Synthetic data from diffusion models improves imagenet classification," *arXiv Preprint*, 2023, arXiv:2304.08466.

9. B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," *arXiv Preprint*, 2023, arXiv:2302.07944.
10. A. Dosovitskiy, N. Houlsby, *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv Preprint*, 2020, arXiv:2010.11929.
11. M. S. Singh, V. Pondenkandath, B. Zhou, P. Lukowicz, and M. Liwicki, "Transforming sensor data to the image domain for deep learning an application to footstep detection," in *Proceedings of International Joint Conference on Neural Networks*, 2017, pp. 2665-2672.
12. L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Transactions on Industrial Electronics*, Vol. 65, 2017, pp. 5990-5998.
13. C. L. Yang, Z. X. Chen, and C. Y. Yang "Sensor classification using convolutional neural network by encoding multivariate time series as two-dimensional colored images," *Sensors*, Vol. 20, 2019, pp. ----.
14. Y. Zhu, M. M. Rahman, and M. A. Ul Alam, "Augmenting deep learning adaptation for wearable sensor data through combined temporal-frequency image encoding," *arXiv Preprint*, 2023, arXiv:2307.00883.
15. H. Rahadian, S. Bandong, A. Widyotriatmo, and E. Joelianto, "Image encoding selection based on Pearson correlation coefficient for time series anomaly detection," *Alexandria Engineering Journal*, Vol. 82, 2023, pp. 304-322.
16. G. R. Garcia, G. Michau, M. Ducoffe, J. S. Gupta, and O. Fink, "Temporal signals to images: Monitoring the condition of industrial assets with deep learning image processing algorithms," *Journal of Risk and Reliability*, Vol. 236, 2022, pp. 617-627.
17. Z. Ahmad and N. Khan, "Towards improved human action recognition using convolutional neural networks and multimodal fusion of depth and inertial sensor data," in *Proceedings of IEEE International Symposium on Multimedia*, 2018, pp. 223-230.
18. P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in Neural Information Processing Systems*, Vol. 34, 2021, pp. 8780-8794.
19. H. J. Kim and N. Moon, "TN-GAN-based pet behavior prediction through multiple-dimension time-series augmentation," *Sensors*, Vol. 23, 2023, pp. ----.
20. A. Vehkaoja, O. Vainio, *et al.*, "Description of movement sensor dataset for dog behavior classification", *Data in Brief*, Vol. 40, 2022, p. 107822.
21. P. Kumpulainen, A. Vehkaoja, *et al.*, "Dog behaviour classification with movement sensors placed on the harness and the collar," *Applied Animal Behavior Science*, Vol. 241, 2021, p. 105393.
22. Z. Wang and T. Oates, "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks," in *Proceedings of Workshops at the 29th AAAI Conference on Artificial Intelligence*, 2015, pp. ----.
23. J. P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *World Scientific Series on Nonlinear Science Series A*, Vol. 16, 1995, pp. 441-446.

Jeong-Hyeon Park received BS degree in Department of Computer Science and Engineering, Hoseo University in 2023. Since March 2023, he is with the Department of Computer Science and Engineering, Hoseo University as an MS candidate. His current research interests big-data processing and analysis.

Nammee Moon received BS, MS, and Ph.D. degrees from the School of Computer Science and Engineering at Ewha Womans University in 1985, 1987, and 1998, respectively. She served as an assistant professor at Ewha Womans University from 1999 to 2003, a 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, HCI and user-centric data, and big data processing and analysis.