Multilabel Classification Model for Infant Activity Recognition Using Single Inertial Sensor

Ayaka Onodera and Riku Ishioka, *The University of Tokyo, Tokyo, 153-8505, Japan*
Yuuki Nishiyama, *The University of Tokyo, Chiba, 277-8568, Japan*
Kaoru Sezaki, *The University of Tokyo, Tokyo, 153-8505, Japan*

Recording and sharing childcare information is crucial for accurately assessing a child's health status and taking appropriate action in case of illness or other emergencies. Although numerous applications and systems have been proposed to assist in recording and sharing these records, the process is still performed manually, presenting a significant burden for parents. Therefore, automatic recording of infants' daily activities is required. In this study, we implement a machine learning model to recognize multilabeled infant activities using a chest-mounted low-sampling rate accelerometer. We collected accelerometer data from 24 h infants between 6 and 24 months as a dataset. Based on the data, we extracted 25 time- and frequency-domain features calculated from the single accelerometer and user features to recognize the 14 daily activities. The performance evaluation considering multilabel classification showed that our proposed model reaches over 88% in the F1 score in the best case.

In childcare, recording and sharing information about daily events, such as sleep schedules, physical activity, and meal times with family members, is essential to the accurate assessment of a child's health status and the appropriate response to emergencies, such as an illness or unforeseen accidents. To support this, several tools (e.g., baby+[a], sprout,[b] and Papatto Childcare[c]) have been widely used to track and share behaviors. Although these applications have partially alleviated the burden of behavior recording, they still require significant manual input, which is still a high burden for busy parents.

In particular, dual-earner households need efficient ways of recording and sharing responsibilities in order to balance work and childcare. According to a survey conducted in the U.S. between 2015 and 2017, 49.9% of households with children under the age of six had both parents working full-time. If cases in which one parent works part-time are included, then the proportion of dual-income households reaches 63.1%.[1] The result of our survey (see the section "Preliminary Investigation: Importance and Burden of Childcare Records") of 200 parents (100 male and 100 female) shows a high demand for the automatic recording of infants' behaviors. Forty respondents had used a childcare record application. Among them, 92% found the applications beneficial, but 80% were dissatisfied with the manual recording process. In addition, manual recording was a major reason many parents did not use childcare recording applications.

Numerous human activity recognition techniques using various sensors have been proposed in pervasive computing and ubiquitous computing, including methods by which to recognize the activities of infants and young children. In studies with infants as subjects, cameras, microphones, and wearable motion

[a][Online]. Available: https://philips-digital.com/baby-new/
[b][Online]. Available: https://sprout-apps.com/sprout-baby-iphone-ipad-app/
[c][Online]. Available: https://papaikuji.info/

1536-1268 © 2024 IEEE
Digital Object Identifier 10.1109/MPRV.2024.3462483
Date of publication 14 October 2024; date of current version 23 January 2025.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.

sensors are commonly used in automatic behavior recording systems in homes and daycare centers. Research has shown that fixed cameras[2,3,4] and microphones[5,6] can be combined with machine learning methods to enable advanced behavior recognition. However, these fixed sensor-based approaches measure activities based on their installation locations, making it impossible to detect activities outside their coverage area. Furthermore, privacy concerns regarding cameras and microphones are major concerns in childcare monitoring.

THEIR PROPOSED SPEECH CLASSIFICATION MODEL ACHIEVED AN ACCURACY OF 97%.

Human behavior recognition technology using wearable sensors, such as accelerometers, gyroscopes, and barometers, has been studied for many years, and its applications in childcare have also been investigated.[5,7,8,9,10] However, most studies have focused on children aged two years and older. Furthermore, these existing studies involve with single-labeled tasks, i.e., tasks in which a ground truth label is associated with a single activity. It has been noted that their recognition performance may be degraded in multilabeled tasks, in which multiple ground truth labels are associated with a single activity.[11] For example, as multilabeled activities, various combinations such as "eating with sitting posture," "sleeping with lying down position," and "drinking milk with holding in a sideways position" would be possible in infant activities.

This study aimed to propose a method by which to use wearable motion sensors to recognize the daily activities of infants between ages 6 and 24 months. Children in this age group are characterized by the dynamic acquisition of movements as they grow older. In addition, it is known that some infant's behaviors may have overlapping behavior labels. Hence, this study develops an activity recognition model that considers age and supports multilabel classification tasks to address these characteristics. The proposed methods are evaluated via cross-validation with 14 labeled motion sensor data collected from 24 infants.

The proposed recognition method for infant daily activities enables a deep and accurate understanding of infant life rhythms and growth in the healthcare and medical domains. Moreover, it can be applied as a fundamental function in the development of various applications for infants and childcare service providers in the pervasive and ubiquitous computing communities.

The main contributions of this study are as follows.

- We conducted a questionnaire survey among childcare providers regarding their attitudes toward behavior recording and the automation of behavior monitoring to investigate methods of recognizing behavior aimed at automating childcare behavior recording.
- We created a sizable childcare dataset comprising 59 h of daily behaviors of 24 subjects below two years of age.
- We developed a method by which to recognize multilabeled infant activities using a chest-mounted low-sampling rate accelerometer. Our evaluation showed that the proposed method achieved an F1-score of 0.88 in the classification of 14 behaviors.

RELATED WORK

Existing research regarding infant activity recognition tasks can be categorized into two major types: fixed-sensor-based[2,5] and wearable-sensor-based[7,8,9,10] approaches. This section summarizes these existing methods and their advantages and disadvantages.

Fixed Sensor-Based Behavior Sensing

Cameras and microphones are typical sensors used in fixed sensor-based approaches. For example, Darapaneni et al.[2] used convolutional neural network (CNN)-based surveillance cameras to detect children's behaviors (standing, sitting, looking, talking, bending and bowing, holding, listening, standing up, and clapping) and emotions in a daycare center. With this method, caregivers can monitor the behavior and emotions of children under their care with 92% accuracy, without constant monitoring. Similarly, García-Domínguez et al.[5] considered environmental sounds to classify four types of child activities (crying, playing, running, and walking). Their proposed speech classification model achieved an accuracy of 97%. Although these methods are highly accurate in action detection, they have significant limitations owing to the installation location. In addition to being unable to record the target behavior outside the installation location, privacy issues due to cameras and sound cannot be ignored. Thus, camera- and voice-based approaches are limited in their applications to behavior recording.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.

Wearable Sensor-Based Behavior Sensing

Wearable devices equipped with accelerometers, gyro sensors, and barometric pressure sensors are widely used for human activity recognition, such as step count, movement mode, and sleep-state detection, beyond adults and children.[7,8,9,10] For example, Nam et al.[7] classified 11 types of daily activities (i.e., wriggling, rolling, standing, standing up, sitting, walking, waddling, crawling, climbing, getting down, and standing still), achieving 98.43% accuracy with ten subjects. However, the studies focused on infants aged 16 to 29 m and did not address early toddlers (0 to 2 years old), which requires childcare records. Moreover, the methods use sensors with high sampling rates (50–75 Hz) and high power consumption, which may not be suitable for long-term behavior recordings using wearable devices. Onodera et al.[11] proposed a method to detect the eight activities of infants and toddlers (6 to 24 m) for childcare recording, including sleeping, crawling, walking, standing, sitting, drinking milk, eating baby food, and holding by a caregiver. Using a low-sampling rate accelerometer sensor and feature extractions, their proposed machine learning (ML) methods classify these behaviors with nearly 80% accuracy. However, it has been noted that classification accuracy is reduced when one label coexists with another (i.e., in multilabel environments), such as false positives for "eating baby food" and "sitting."

Single- and Multilabel Classification

Multilabel classification is a classification problem where an instance (such as an object, event, or action) is simultaneously associated with multiple labels. Unlike single-label classification, where each instance is assigned only one label, multilabel classification allows for detecting and assigning multiple relevant labels to a single instance. Conventional ML methods such as support vector machine (SVM), k-nearest neighbor (KNN), decision tree (DT), and random forest (RF) are commonly used to address the single-label classification task. These conventional ML methods such as SVM, KNN, DT, and RF are commonly used to address the single-label classification task. These traditional ML methods require minimal data and computational power[12] and are characterized by greater interpretability than state-of-the-art deep learning-based methods. ML methods require minimal data and computational power[12] and are characterized by greater interpretability than state-of-the-art deep learning-based methods.

However, simple single-label classification can decrease classification performance in scenarios where labels like "riding a train" and "sitting" coexist. Multilabel classification is required to address such multilabel tasks. Binary relevance (BR), label powerset (LP), and classifier chain (CC)[13] are traditional ML models for multilabel classification tasks. BR is the simplest multilabel classification method and is used as the baseline method for multilabel classification tasks. This method performs multilabel classification by decomposing each label into an independent binary classification problem. BR decomposes each label into an individual binary classification problem and trains independent classifiers. It is easy to implement and works well when labels are assumed to be independent but may result in lower accuracy by ignoring label correlations. LP treats all label combinations as new classes, considering label correlations, but can increase computational costs with many label combinations. CC orders labels and trains classifiers sequentially, using previous label predictions as features. This method often achieves high performance but is influenced by label order.

For instance, multilabel classification methods have been applied in various kinds of research regarding human-behavior recognition.[14,15,16,17] Jethanandani et al.[14] proposed a BR-based activity recognition method to address multiresident activity recognition in the smartphone environment with its embedded sensors.

PRELIMINARY INVESTIGATION: IMPORTANCE AND BURDEN OF CHILDCARE RECORDS

As a preliminary investigation, a survey was conducted on children's awareness of behavior recording and people who take care of their children regarding the automation of behavior recording to examine behavior recognition methods aimed at automating the recording of childcare behaviors.

This preliminary investigation conducted a questionnaire survey on the awareness of behavioral records of infants and the necessity of automating such records. The study surveyed people with experience in child-rearing. Furthermore, we sought to improve user convenience through the automation of behavioral records. The survey was conducted using the survey implementation service CrowdWorks.[d] The survey's content was analyzed to

[d][Online]. Available: https://crowdworks.jp/

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.

understand the need for behavior records and develop a more convenient behavior record service for infants and toddlers. Questionnaire content analysis was conducted to determine the demand for behavior records and to develop more convenient infant activity recording services. The survey targeted 200 participants, 100 men, and 100 women, aged 18 years or older, with experience in child-rearing. Participants registered with CrowdWorks, a web survey service, were recruited.

The survey results showed that 90% of respondents (200 people with child-rearing experience) were in their 20s–40s, reflecting the views of households with active children. Participants with behavior recording experience often share their children's daily activities with their cohabiting partners. Ninety percent of those with behavior recording experience perceived that behavior recording has benefits, implying that they were more likely to benefit from understanding and sharing the children's daily rhythm and growth rather than their physical condition or development. However, 80% indicated that behavior recording is inconvenient, which they attributed to the burden of recording, forgetting to record, or being unable to record. Over 95% of those with experience in behavior recording stated that automated recording would increase convenience. Most participants expected that this would reduce the time spent recording. Even those without experience in behavior recording indicated that the recording was burdensome, and approximately 90% of the respondents expressed expectations for automated recording. Participants without recording experience were most likely to anticipate automation to enable them to track their growth process. Those with experience in behavior recording were more interested in participating in the children's daily life, whereas those without experience in behavior recording were more interested in identifying abnormalities than in understanding the daily life of the children.

MOTIVATION AND APPROACH

This study aimed to automatically detect multiple behaviors in infants from 6 to 24 months. The questionnaire-based preliminary investigation has shown numerous advantages to automatically recording infant activities, such as monitoring infants' health, determining their state of growth, and sharing their activities with others. However, owing to the significant developmental changes in infants from 6 to 24 months, it is essential to construct models considering the age in months. Decreased detection accuracy owing to overlapping behavior labels is also a challenge. Sections "Overlapped Behavior Label Issue" and "Behavioral Development Across Growth Stages" describe more detailed information about the challenges and approaches to addressing the challenges in detail.

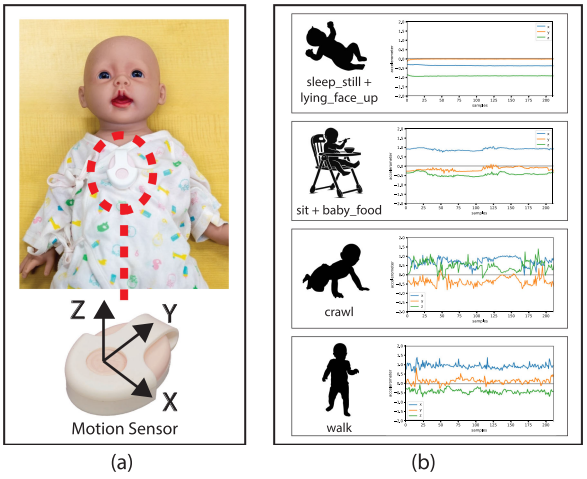

FIGURE 1. Mounting positions of motion sensor and examples of infant behaviors. (a) Sensor setup. (b) Example of infant's behavior.

To achieve this goal, we first defined the target activities of infants. Second, based on the above definitions, behavior data were collected from the target participants (N = 24) using a wearable device equipped with a low sampling rate motion sensor, as shown in Figure 1(a), and its features were analyzed. Finally, we developed, evaluated, and discussed machine learning models for multilabel classification. This experimental process was conducted with the approval of the Experimental Ethics Review Committee of the University of Tokyo.

Overlapped Behavior Label Issue

Infant behavior often comprises multiple concurrent actions, potentially reducing the accuracy of behavior recognition.[11] For instance, the action of "eating baby food" includes the action of "sitting," making it difficult to accurately classify using simple multilabel classification methods, as shown in Figure 1(b). Other examples include sleep, breastfeeding, and bottle-feeding, which involve various postures (e.g., lying face down, lying face up, lying on the side). To accurately detect these behaviors, it is necessary to employ methods capable of detecting multiple labels simultaneously.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.

TABLE 1. Infants behavior dataset.

Large class	Subclass	Data size (minutes)	Provider (N)	Age of month (mean/std)
stay	lying_face_up	846.03	16	11.44 / 3.60
	lying_face_side	909.95	16	11.94 / 3.60
	lying_face_down	625.11	13	12.85 / 3.21
	sit	112.14	24	12.17 / 3.85
	stand	27.12	15	14.20 / 3.30
move	crawl	10.82	16	12.44 / 3.08
	walk	9.39	8	16.50 / 2.51
sleep	sleep_still	2381.09	22	11.95 / 3.88
	sleep_fidgety	74.05	22	12.09 / 4.05
meal	milk	68.33	15	12.80 / 4.21
	baby_food	131.90	21	12.24 / 3.95
hold	hold_horizontal	102.59	18	11.61 / 3.99
	hold_vertical	158.63	24	12.04 / 4.00
	piggyback	91.98	15	12.13 / 4.14

Behavioral Development Across Growth Stages

Infants acquire various movements and postures according to their developmental stages. Especially up to around 2 years old, their motor skills develop rapidly, and the postures and movements they can perform differ by month. For example, they acquire movements in the following order: supine position (lying on their back), prone position (lying on their stomach), sitting position, standing position, and walking. In addition, the timing of acquiring these movements varies among individuals. According to a report by the World Health Organization,[18] 1% of infants can walk independently at 8.2 months, 50% at 12 months, and 99% at 17.6 months. This indicates that using age information can potentially give an understanding of an infant's developmental progress to some extent.

MULTILABEL CLASSIFICATION MODEL FOR INFANTS' BEHAVIORS

This section describes designing and implementing a multilabel classification model for classifying infants' behaviors.

Definition of Target Behaviors

We defined the large class as typical movements commonly performed by infants and the subclass as movements they become capable of performing during growth. The large class contains five classes: stay, move, hold, meal, and sleep, which are based on the definition of previous work.[11] During data collection with actual infants, as described in the section "Dataset," parents were requested to provide motion sensors and video recordings of the behaviors associated with the defined large class. Based on the provided data, we organized the typical behaviors performed by children at each developmental stage into subclasses. As a result, the subclass contains the following 14 labels, which are presented in Table 1.

Dataset

The infant motion dataset was collected using ICUCO,[e] a wearable chest-worn device designed for infants as illustrated on Figure 1(a). The wearable device can collect skin temperature, heart rate, and three-axis accelerometers, and transfer these to a connected smartphone via Bluetooth. The data collection process is followed by a prior work.[11] Moreover, this device is widely used in many nurseries to detect sleep patterns, particularly in the prone position in Japan.

The target behaviors—which contain five large classes and 14 subclasses—are listed in Table 1. Twenty-four infants (nine males and 15 females), averaging 12.0 (std: 4.0, min: 6, mid: 11, max: 20) m of age, participated in the data collection experiment. While wearing a wearable

[e][Online]. Available: https://www.icuco.co.jp/

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.

device, infants performed the large class behaviors that were recorded using a time-stamped video camera. Data were collected only for target behaviors that could be performed as the infants grew. Sensor data were acquired at intervals of 150 ms (approximately 7 Hz). After data collection, the time-stamped videos were reviewed and labeled with the subclass.

Feature Extraction

This research used 25 features extracted from an accelerometer sensor and user attributes. We extracted time- and frequency-domain features from each acceleration axis (*x*, *y*, and *z*) and the triaxial composite acceleration. A sliding window was used for feature sampling, capturing 16 samples (approximately 2 s) at a time and moving forward by 8 samples after each extraction. This window size has reached the best estimation performance in a previous similar behavior detection task.[11] The time-domain features include mean, standard deviation (std), skewness (skew), regression coefficient (coef), and kurtosis (kurt) of accelerometer data from each axis and its composition. Moreover, the frequency-domain features contain entropy from each axis and its composition. In addition, each participant's age of month was added as a feature to represent their growth process. The extracted features are then subjected to *z*-normalization.

Figure 2 shows the Pearson correlation between each feature. The features used in the previous paper, such as maximum and minimum acceleration, median, and root mean square (RMS), highly correlated with other features—more than 0.8. Features with high correlations were excluded from the features used in this study because of the risk of overfitting.

The dataset used in this study includes both actions that infants can perform for a long time and actions that they cannot perform, resulting in an imbalanced dataset. Therefore, undersampling was performed to mitigate the data size imbalance. Specifically, 100 samples were extracted for each user and action, and these were used for the training and evaluation of the algorithm. This indicates that the maximum number of samples per action is 2400. Random sampling was utilized as the sampling method. Actions that could not be performed were treated as missing values and were not used in this training.

Multilabel Classification Method

As a multilabel classification method, we used XGBoostring (XGB), random forest (RF), *k*-nearest neighbor algorithm (KNN), multilayer perceptron regressor

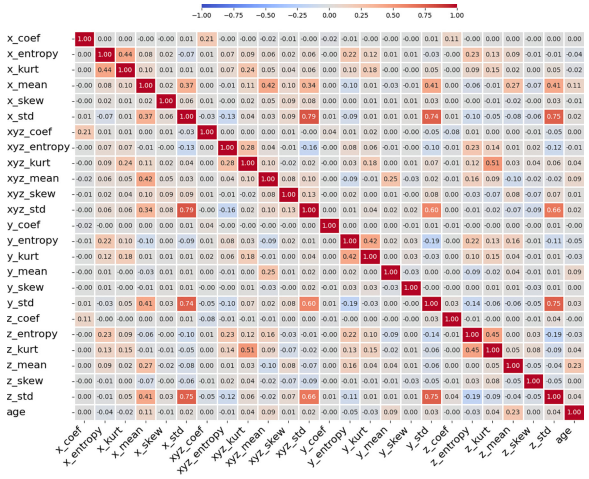

FIGURE 2. Pearson correlation between each extracted feature from accelerometer and age information.

(NN), extra tree (ET), BR, LP, and CC to classify multilabeled behaviors. Moreover, BR, LP, and CC use the following binary classification models: LGBM, Linear Regression (LR), RF, and ET. These classification models are implemented using scikit-multilearn[19] and scikit-learn,[f] which is a Python library. The Python, scikit-multilearn, and scikit-learn versions are 3.10.0, 0.2.0, and 1.4.2, respectively. Moreover, the learning and evaluation process is on Ubuntu 20.04.6 with AMD Ryzen Threadripper 3970X CPU. RF, LR, KNN, NN, and ET are integrated into scikit-learn. XGB[g](version 2.1.0) and LGBM[h](version 4.4.0)[20] are installed via a package manager on Python. The hyperparameters for each multilabel and internal classification model are used as the default settings of each library.

EVALUATION

The performance of the proposed model was evaluated using the collected dataset in the following steps. First, to compare the performance of the conventional multilabeled classification models, we evaluated the performance differences between 14 multilabel classifiers (see the sections "Comparison Across Multilabel Classification Methods"). Second, the combination of the top-performing multilabeled classifiers and their internal binary classifiers is evaluated on untrained validation data to

[f][Online]. Available: http://scikit.ml/index.html
[g][Online]. Available: https://xgboost.readthedocs.io/en/stable/
[h][Online]. Available: https://lightgbm.readthedocs.io/en/stable/

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.

TABLE 2. Performance comparison of multilabel classification models in four-hold cross validation.

Model	Accuracy	Precision	Recall	F1-score
LP+LGBM	**0.79**	0.85	**0.86**	**0.86**
LP+RF	0.77	0.85	0.84	0.84
LP+ET	0.78	0.86	0.85	0.85
LP+LR	0.49	0.66	0.65	0.64
CC+LGBM	0.76	0.84	0.84	0.83
CC+RF	0.69	0.85	0.78	0.80
CC+ET	0.70	0.84	0.79	0.79
CC+LR	0.35	0.61	0.56	0.54
BR+LGBM	0.67	0.89	0.79	0.83
BR+RF	0.59	0.92	0.71	0.78
BR+ET	0.56	0.94	0.68	0.75
BR+LR	0.17	0.62	0.47	0.50
XGB	0.68	0.89	0.80	0.84
RF	0.58	0.93	0.70	0.77
ET	0.58	**0.95**	0.70	0.77
NN	0.50	0.77	0.69	0.71
KNN	0.37	0.68	0.53	0.56

determine the classification performance for each label (see the section "Classification Performance of Each Behavior"). Finally, subject-specific leave-one-out cross-validation (LOOCV) is conducted to assess its applicability to unlearned subjects (see the section "Performance Generality").

The dataset was split into training and validation subsets with a split ratio of 80 and 20, respectively. Moreover, this evaluation measures accuracy, recall, precision, and F1-score as a performance of each model. Each sore is calculated using metrics[i] library in scikit-learn. Because this classification task involves multilabel classification, the scores tend to be lower than those for single-label classification. Accuracy tends to be especially low because it is only considered a true positive or true negative when all labels (16 in this case) are correctly predicted.

Comparison Across Multilabel Classification Methods

In this overall performance comparison, we compared the performance of the proposed method using 17 multilabel classifiers. The classifiers contain eight classification models: LP, CC, BR, XGB, RF, ET, KNN, and NN. In addition, four internal classifiers (ET, LGBM, RF, and LR) are used on LP, CC, and BR. Table 2 presents each multilabel classification model's accuracy, precision, recall, and F1 score through four-hold cross-validation using the datasets. The cross-validation was performed with three-quarters of the training subset as training data and one-quarter as test data.

Overall, the LP-based models also performed better than those using other classifiers, regardless of the internal classifiers. The LP+LGBM model showed the highest performance, with an F1 score of 0.86, across the 17 multilabel classifiers. XGB, which can output multilabels directly, also shows a relatively high F1 score of 0.84, but there is a discrepancy between precision 0.89 and recall 0.80. Overall, LP+LGBM is more suitable for this classification task than XGB.

TABLE 3. Classification performance of the LP+LGBM on the validation data.

	Precision	Recall	F1-score	Support
sleep_still	0.99	0.97	0.98	1290
lying_face_side	0.97	0.97	0.97	586
lying_face_up	0.97	0.97	0.97	597
lying_face_down	0.97	0.96	0.97	417
hold_horizontal	0.88	0.93	0.91	400
milk	0.84	0.89	0.86	305
crawl	0.80	0.91	0.85	77
sleep_fidgety	0.82	0.88	0.85	269
sit	0.80	0.83	0.82	797
baby_food	0.73	0.82	0.77	466
hold_vertical	0.76	0.72	0.74	456
piggyback	0.74	0.69	0.71	265
stand	0.70	0.66	0.68	196
walk	0.66	0.62	0.64	71
weighted avg	0.88	0.89	0.88	6192

Classification Performance of Each Behavior

Based on the comparison across multilabel classification methods, we conducted a performance evaluation of the LP+LGBM model, which demonstrated superior performance, using the validation data. Moreover, in this evaluation, we used a grid search method to optimize the hyperparameters of LP+LGBM. The

[i][Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.

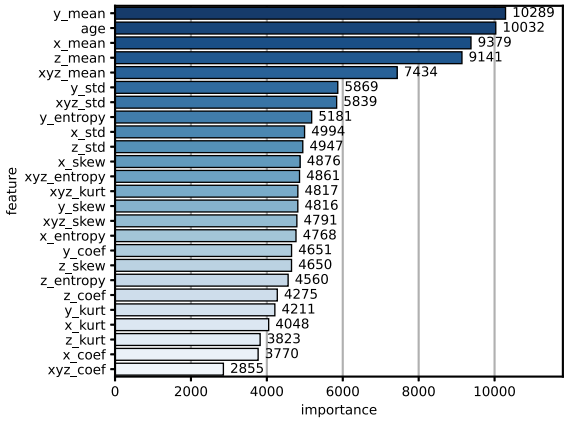

FIGURE 3. Feature importance of LP+LGBM model.

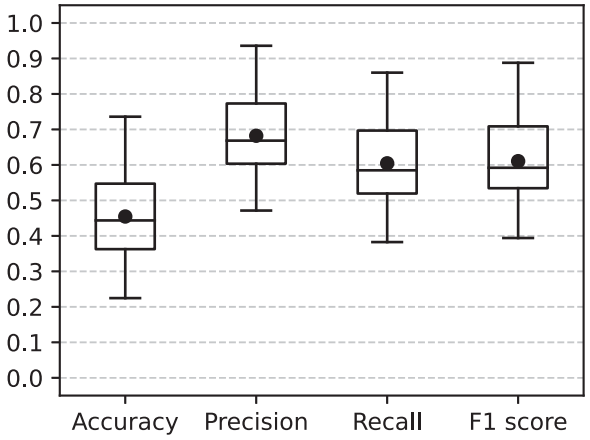

FIGURE 4. Classification performance per infant using the LP +LGBM model.

hyperparameters were L1 regularization, L2 regularization, maximum depth, number of leaves, and learning rate set to 0.5, 0.0, 20, 100, and 0.05, respectively. This configuration maximized the model's performance.

Table 3 shows the results of the LP+LGBM model with each target behavior. Overall, the precision, recall, and F1-score were 0.88, 0.89, and 0.89, respectively, demonstrating performance nearly equivalent to that observed during cross-validation. When evaluated on individual activity labels, particularly for horizontal activities such as `sleep_still` and `lying` behaviors, the classification accuracy was notably high, with F1-scores exceeding 0.97. However, there is room for improvement in classification performance for vertical behavior-contained activities. For instance, activities such as `walk`, `stand`, `piggyback`, and `hold_vertical`, and have less than 0.75 F1 scores. Recall values of these lower performance activities tend to be significantly lower than the precision values.

Feature Importance Analytics

Figure 3 shows the importance of 25 features on the LP+LGBM model with a box plot. This figure demonstrates the strong influence of the standard deviation and means of each axis of the accelerometer, as well as the age in months, in the estimation of infant behavior. Other accelerometers calculated from the accelerometers, such as kurtosis, coef, skew, and entropy, also contribute to the estimation of behavior, although they are relatively low. Including the feature "age in months" in the tree-based classifier (i.e., LightGBM) suggests that it may be possible to separate doable and undoable behaviors by age in months.

Performance Generality

To evaluate the generalization performance of the proposed model, we conducted subject-based LOOCV. The LP+LGBM was trained on data from 23 individuals and evaluated on data from one individual; as LOOCV, the loop was repeated 24 times with different subject combinations.

Figure 4 illustrates the result, including accuracy, precision, recall, and F1 score as a box plot. As shown in Figure 4, the F1 score, recall, and precision performed less than the result of cross-validation explained in the "Comparison Across Multilabel Classification Methods" section. The mean and median F1 scores were 0.61 and 0.59, respectively. However, the classification performance varied between individuals, with the highest F1 scores reaching 0.89 while the lowest F1 scores were below 0.39.

DISCUSSION

In this study, we demonstrated that 14 types of activities with an F1-Score of 0.88 can be classified using only simple accelerometer sensors and age labels. In particular, the LP+LGBM model showed high performance, and including age as a feature allows for activity recognition across a wide range of age groups. Conversely, using multilabel classification can address the issue of overlapping labels, it necessitates estimating 14 types of activities for a single action, which tends to improves precision but decreases recall.

In addition, the current features resulted in relatively low detection accuracy for vertical motion movements. Therefore, it is necessary to explore

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.

methods such as using different features like peak detection, widening the sliding window, or employing algorithms that can utilize deep learning to learn the waveform of accelerometer data. For example, with `piggyback` or `hold`, vertical periodic oscillations caused by the caregiver's "walking" or "rocking" were observed. Thus, these features may be effective for vertical motion recognition. Furthermore, LOOCV results indicated significant variability in classification performance among individuals, suggesting the need for personalization mechanisms in addition to the basic classification model proposed.

Children's behavior patterns in childcare environments exhibit a certain degree of periodicity. Thus, leveraging "time information" and "preceding and subsequent activity data" could enhance effectiveness.

This study used 24 infants, and the sensor data used were limited to accelerometer data. Therefore, future challenges include studies involving more participants and various sensors. In addition, the application and performance evaluation of the proposed method in real-life environments with infants remain future tasks.

CONCLUSION

This study proposes a method for classifying 14 types of infant activities that can simultaneously overlap the same action. For this purpose, we employed a chest-mounted low-sampling rate accelerometer commonly used in childcare settings and developed a multilabel infant activity classification model using the sensor data from the device. As a dataset of the infant activities, we collected triaxial accelerometer data from 24 infants aged 6–24 months. Based on the dataset, the proposed method extracts 25 features related to the user, time, and frequency domains, and a multilabel classification detects each infant's activity. Consequently, the proposed method achieved an F1-score of approximately 0.88 in LP+LGBM. Although horizontal movements (e.g., `sleep_still`, `hold_horizontal`, and `crawl`) showed high recognition accuracy, the classification accuracy of movements with a vertical movement component (e.g., `walk`, `piggyback`, and `stand`) tended to be low. Therefore, in addition to the current features, it was found necessary to add features that reflect vertical movements to improve accuracy. The classification performance might be improved by incorporating features that consider longer periods, the number of peaks of accelerometer, contexts before and after, and time features. This study provides a baseline for detecting multiple activities in infants.

ACKNOWLEDGMENTS

This work was supported in part by First-Ascent Inc., in conjunction with a JSPS KAKENHI under Grant 23K17004.

The author's Ayaka Onodera and Riku Ishioka are contribution to this work was made during their master's program, and they are currently affiliated with industry.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Research Ethics Committee of the University of Tokyo under Application No. 22-129.

REFERENCES

1. J. Sullivan, "Comparing characteristics and selected expenditures of dual-and single-income households with children," *Monthly Lab. Rev.*, vol. 143, pp. 1–14, 2020.
2. N. Darapaneni et al., "Activity & emotion detection of recognized kids in CCTV video for day care using SlowFast & CNN," in *Proc. IEEE World AI IoT Congr.*, 2021, pp. 0268–0274.
3. P. Wei, D. Ahmedt-Aristizabal, H. Gammulle, S. Denman, and M. A. Armin, "Vision-based activity recognition in children with autism-related behaviors," *Heliyon*, vol. 9, no. 6, 2023, Art. no. e16763.
4. S. Suzuki, Y. Amemiya, and M. Sato, "Skeleton-based explainable human activity recognition for child gross-motor assessment," in *Proc. 46th Annu. Conf. IEEE Ind. Electron. Soc.*, 2020, pp. 4015–4022.
5. A. García-Domínguez et al., "Children's activity classification for domestic risk scenarios using environmental sound and a Bayesian network," *Healthcare*, vol. 9, no. 7, 2021, Art. no. 884.
6. G. Laput, K. Ahuja, M. Goel, and C. Harrison, "Ubicoustics: Plug-and-play acoustic activity recognition," in *Proc. 31st Annu. ACM Symp. User Interface Softw. Technol.*, New York, NY, USA: Association for Computing Machinery, 2018, pp. 213–224. [Online]. Available: https://doi.org/10.1145/3242587.3242609
7. Y. Nam and J. W. Park, "Child activity recognition based on cooperative fusion model of a triaxial accelerometer and a barometric pressure sensor," *IEEE J. Biomed. Health Inform.*, vol. 17, no. 2, pp. 420–426, Mar. 2013.
8. S. Kwon, P. Zavos, K. Nickele, A. Sugianto, and M. V. Albert, "Hip and wrist-worn accelerometer data analysis for toddler activities," *Int. J. Environ. Res. Public Health*, vol. 16, no. 14, 2019, Art. no. 2598. [Online]. Available: https://www.mdpi.com/1660-4601/16/14/2598

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.

9. S. G. Trost, D. P. Cliff, M. N. Ahmadi, N. V. Tuc, and M. Hagenbuchner, "Sensor-enabled activity class recognition in preschoolers: Hip versus wrist data," *Med. Sci. Sports Exercise*, vol. 50, pp. 634–641, 2017.
10. S. Kurashima and S. Suzuki, "Improvement of child activity recognition algorithm for accurate calculation of consumption calorie," in *Proc. 42nd Annu. Conf. IEEE Ind. Electron. Soc.*, 2016, pp. 5915–5920.
11. A. Onodera, R. Ishioka, Y. Nishiyama, and K. Sezaki, "Assessing infant and toddler behaviors through wearable inertial sensors: A preliminary investigation," in *Proc. Companion Publication 25th Int. Conf. Multimodal Interact.*, 2023, pp. 16–20.
12. F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey," *IEEE Access*, vol. 8, pp. 210816–210836, 2020.
13. J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Machine Learning and Knowledge Discovery in Databases*, W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor, Eds. Berlin, Germany: Springer, 2009, pp. 254–269.
14. M. Jethanandani, T. Perumal, Y.-C. Liaw, J.-R. Chang, A. Sharma, and Y. Bao, "Binary relevance model for activity recognition in home environment using ambient sensors," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2019, pp. 1–2.
15. M. Jethanandani, T. Perumal, J.-R. Chang, A. Sharma, and Y. Bao, "Multi-resident activity recognition using multi-label classification in ambient sensing smart homes," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2019, pp. 1–2.
16. R. Kumar, I. Qamar, J. S. Virdi, and N. C. Krishnan, "Multi-label learning for activity recognition," in *Proc. Int. Conf. Intell. Environ.*, 2015, pp. 152–155.
17. R. Mohamed, T. Perumal, M. Sulaiman, and N. Mustapha, "Multi-resident activity recognition using label combination approach in smart home environment," in *Proc. IEEE Int. Symp. Consum. Electron.*, 2017, pp. 69–71.
18. W. M. G. R. S. GROUP and M. de Onis, "WHO motor development study: Windows of achievement for six gross motor development milestones," *Acta Paediatrica*, vol. 95, no. S450, pp. 86–95, 2006.
19. P. Szymański and T. Kajdanowicz, "Scikit-multilearn: a scikit-based Python environment for performing multi-label classification," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 209–230, 2019.
20. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30. Red Hook, NY, USA: Curran Associates, Inc., 2017.

AYAKA ONODERA is a computer science researcher with the Institute of Industrial Science, The University of Tokyo, Tokyo, 153-8505, Japan. Her research interests focus on activity recognition and wearable sensing. Onodera received her master's degree in socio-cultural environmental studies from the University of Tokyo. Contact her at ndr.yk0909@mcl.iis.u-tokyo.ac.jp.

RIKU ISHIOKA is a computer science researcher with the Institute of Industrial Science, The University of Tokyo, Tokyo, 153-8505, Japan. His research addresses challenges in environmental sustainability and healthcare through machine learning, data analysis, and sensor-based systems. Ishioka received his master's degree in information science and technology from the University of Tokyo. Contact him at riku.ishioka@mcl.iis.u-tokyo.ac.jp.

YUUKI NISHIYAMA is a lecturer at the Center for Spatial Information Science, The University of Tokyo, Chiba, 277-8568, Japan. His research interests include ubiquitous computing, mobile-wearable sensing, and human-computer interaction. Nishiyama received his Ph.D. degree in media and governance from Keio University, Kanagawa, Japan, in 2017. He is a member of ACM, IEEE, and Information Processing Society of Japan (IPSJ). He is a member of IEEE and corresponding author of this article. Contact him at nishiyama@csis.u-tokyo.ac.jp.

KAORU SEZAKI is a professor of the Institute of Industrial Science, The University of Tokyo, 153-8505, Tokyo, Japan and also with the Center for Spatial Information Science, The University of Tokyo, Chiba, Japan. His research interests include e-Health, sensor networks, IoT, and urban computing. Sezaki received his Ph.D. degree from the University of Tokyo, Tokyo, Japan. He is a member of IEEE. Contact him at sezaki@iis.u-tokyo.ac.jp.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DE MINAS GERAIS. Downloaded on April 02,2025 at 15:52:05 UTC from IEEE Xplore. Restrictions apply.