

Article

Deep Learning Classification of Canine Behavior Using a Single Collar-Mounted Accelerometer: Real-World Validation

Robert D. Chambers ¹, Nathanael C. Yoder ¹, Aletha B. Carson ¹, Christian Junge ¹, David E. Allen ¹, Laura M. Prescott ¹, Sophie Bradley ², Garrett Wymore ¹, Kevin Lloyd ¹ and Scott Lyle ^{1,*}

¹ Pet Insight Project, Kinship, San Francisco, CA 94103, USA; rdchambers@gmail.com (R.D.C.); nyoder@gmail.com (N.C.Y.); aletha.carson@effem.com (A.B.C.); tianjunge@gmail.com (C.J.); david@petinsight.com (D.E.A.); laura.prescott@whistle.com (L.M.P.); garrett.a.wymore.13@gmail.com (G.W.); krcloyd@gmail.com (K.L.)

² WALTHAM Petcare Science Institute, Melton Mowbray, Leicestershire LE14 4RT, UK; Sophie.Bradley@effem.com

* Correspondence: scottlyle@whistle.com

Simple Summary: Collar-mounted activity monitors using battery-powered accelerometers can continuously and accurately analyze specific canine behaviors and activity levels. These include normal behaviors and those that are indicators of disease conditions such as scratching, inappetence, excessive weight, or osteoarthritis. Algorithms used to analyze activity data are validated by video recordings of specific canine behaviors, which were used to label accelerometer data. The study described here was noteworthy for the large volume of data collected from more than 2500 dogs in clinical and real-world home settings. The accelerometer data were analyzed by a machine learning methodology, whereby algorithms were continually updated as additional data were acquired. The study determined that algorithms from the accelerometer data detected eating and drinking behaviors with a high degree of accuracy. Accurate detection of other behaviors such as licking, petting, rubbing, scratching, and sniffing was also demonstrated. The study confirmed that activity monitors using validated algorithms can accurately detect important health-related canine behaviors via a collar-mounted accelerometer. The validated algorithms have widespread practical benefits when used in commercially available canine activity monitors.

Abstract: Collar-mounted canine activity monitors can use accelerometer data to estimate dog activity levels, step counts, and distance traveled. With recent advances in machine learning and embedded computing, much more nuanced and accurate behavior classification has become possible, giving these affordable consumer devices the potential to improve the efficiency and effectiveness of pet healthcare. Here, we describe a novel deep learning algorithm that classifies dog behavior at sub-second resolution using commercial pet activity monitors. We built machine learning training databases from more than 5000 videos of more than 2500 dogs and ran the algorithms in production on more than 11 million days of device data. We then surveyed project participants representing 10,550 dogs, which provided 163,110 event responses to validate real-world detection of eating and drinking behavior. The resultant algorithm displayed a sensitivity and specificity for detecting drinking behavior (0.949 and 0.999, respectively) and eating behavior (0.988, 0.983). We also demonstrated detection of licking (0.772, 0.990), petting (0.305, 0.991), rubbing (0.729, 0.996), scratching (0.870, 0.997), and sniffing (0.610, 0.968). We show that the devices' position on the collar had no measurable impact on performance. In production, users reported a true positive rate of 95.3% for eating (among 1514 users), and of 94.9% for drinking (among 1491 users). The study demonstrates the accurate detection of important health-related canine behaviors using a collar-mounted accelerometer. We trained and validated our algorithms on a large and realistic training dataset, and we assessed and confirmed accuracy in production via user validation.

Keywords: canine; accelerometer; deep learning; behavior; activity monitor



Citation: Chambers, R.D.; Yoder, N.C.; Carson, A.B.; Junge, C.; Allen, D.E.; Prescott, L.M.; Bradley, S.; Wymore, G.; Lloyd, K.; Lyle, S. Deep Learning Classification of Canine Behavior Using a Single Collar-Mounted Accelerometer: Real-World Validation. *Animals* **2021**, *11*, 1549. <https://doi.org/10.3390/ani11061549>

Academic Editor: Lynette A. Hart

Received: 8 April 2021

Accepted: 19 May 2021

Published: 25 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Much as recent progress in smartwatches has enabled new telehealth applications [1–4], recent progress in internet-connected pet wearables, such as collar-mounted activity monitors, has prompted interest in using these devices to improve the cost and efficacy of veterinary care [5]. Just as with smartwatches in human telehealth, accelerometer-based activity monitors have emerged as an inexpensive, low-power, and information-rich approach to pet health monitoring [6–8].

Accelerometer-based pet activity monitors analyze the moment-to-moment movement measured by a battery-powered accelerometer. They are typically attached to the pet via a collar, though attachment methods may be more elaborate in research settings. Using the device's accelerometer signal (sometimes in combination with gyroscope, magnetometer, GPS, or other sensor signals), collar-mounted activity monitors can accurately estimate pet activity levels [9–14], step count, and distance traveled [12].

In recent years, advances in machine learning have allowed pet activity monitors to move beyond estimating aggregate activity amounts, to detecting when and for how long a pet performs common activities such as: walking, running, lying down, or resting [15–17], these biometric capabilities have progressed to include increasingly specific and varied activities such as drinking, eating, scratching, and head-shaking [8,16,18–21].

The benefits of accurate and quantitative behavior detection in pet health are extensive. Pet activity monitors have been shown to be useful in the detection and diagnosis of pruritis [22,23] and in potential early prediction of obesity [24]. They have also been used in monitoring response to treatments such as chemotherapy [25]. Furthermore, statistical analysis of activity and behavior monitoring on large numbers of pets can be an expedient approach to medical and demographic studies [24].

Although several studies have demonstrated and measured the accuracy of activity recognition algorithms [8,16,18–21], the datasets used to train and evaluate the algorithms are typically not representative of the broad range of challenging environments in which commercial pet activity monitors must function. For instance, most existing studies use exclusively healthy dogs and are often run in controlled environments that promote well-defined and easily detectable behaviors with a low risk of confounding activities.

Unfortunately, real-world algorithm performance often lags far behind the performance measured in controlled environments [26,27]. For instance, existing studies typically ensure careful installation of the device in a specific position on a properly adjusted collar. In real-world usage, collars vary in tightness and often rotate to arbitrary positions unless the activity monitor device is very heavy. Collar rotation and tightness [28], as well as the use of collar-attached leashes [29], can compromise performance. In our experience, confounding activities like riding in a car or playing with other pets can produce anomalous results if not adequately represented in training datasets. Finally, some studies use multiple accelerometers or harness-mounted devices [30], which limit applicability in many consumer settings.

The work described here was performed as part of the Pet Insight (PI) Project [31], a large pet health study to enable commercial pet activity monitors to better measure and predict changes in a pet's health by:

- Sourcing training data from project participants and external collaborators to build machine learning training databases and behavior detection models such as those described in this work.
- Combining activity data, electronic medical records, and feedback from more than 69,000 devices distributed to participants over 2–3 years to develop and validate proactive health tools.
- Using the resulting datasets, currently covering over 11 million days in dogs' lives, to enable insights that support pet wellness and improve veterinary care.

This work presents the results of the PI Project's efforts to develop and validate these behavior classification models [32], including an evaluation of model performance in a

real-world context and addressing limitations from controlled research settings such as device fit and orientation.

2. Materials and Methods

2.1. Activity Monitor

Data were collected primarily via a lightweight canine activity monitor (Figure 1, Whistle FIT[®], Mars Petcare, McLean, VA, USA), which was designed and produced specifically for this study. Smaller amounts of data were collected via the commercially available Whistle 3[®] and Whistle GO[®] canine activity monitors. All three devices used the same accelerometer. Unlike the Whistle FIT[®], these latter devices are furnished with GPS receivers and cellular radios. However, in all cases, the behavior classification in this study is performed using only the output of the devices' 3-axis accelerometers.

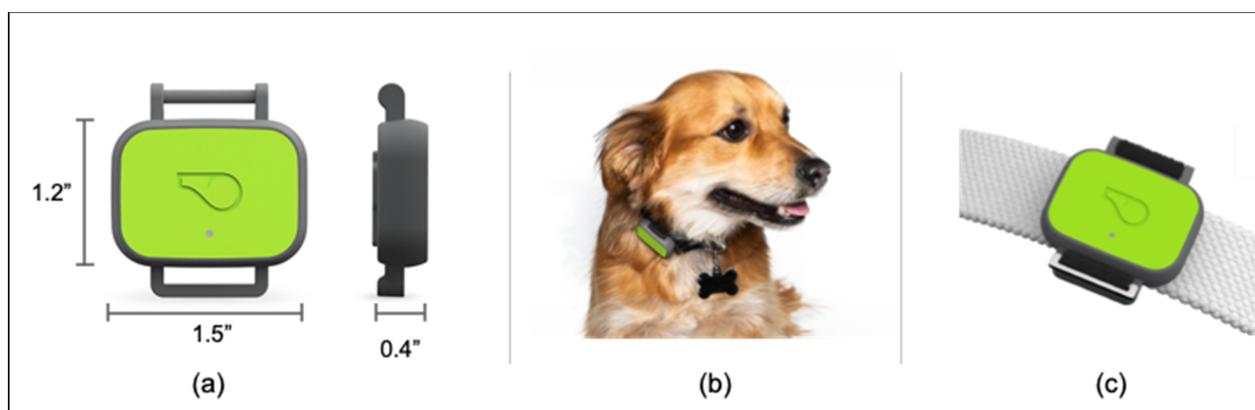


Figure 1. Activity monitors used in this study. Most data in this study were acquired from Whistle FIT[®] activity monitors. Device dimensions are shown in (a), and a device in use is shown in (b). The device often rotates to different positions around each dog's collar. The device can attach to most dog collars up to 1" (25 mm). Attachment detail is shown in (c). The two other devices used this study (the Whistle 3[®] and the Whistle GO[®]) are larger and heavier.

2.2. Accelerometry Data Collection

All monitoring devices acquired accelerometry data and uploaded it according to their usual operation. That is, the devices acquired 25–50 Hz 3-axis accelerometry data for at least several seconds whenever significant movement was detected. Data were compressed and annotated with timing data using a proprietary algorithm. Data were temporarily stored on-device and then uploaded at regular intervals when the devices were in Wi-Fi range. Uploads were processed, cataloged, and stored in cloud-hosted database services by Whistle servers. The compressed accelerometry data were retrieved on demand from the cloud database services in order to create the training, validation, and testing databases used in this study.

2.3. Animal Behavior Data Collection

Animal behavior data were collected (summarized in Table 1 and further described elsewhere in this report) and used to create two datasets used in model training and evaluation:

- *Crowd-sourced (crowd) dataset.* This dataset contained both (a) long (multi-hour) in-clinic recordings, as well as (b) shorter recordings submitted by project participants. This large and diverse dataset was meant to reflect real-world usage as accurately as possible.
- *Eating and drinking (eat/drink) dataset.* This dataset consisted of research grade sensor and data using a protocol designed to represent EAT and DRINK behaviors. Other observed behaviors were incidental.

Table 1. Datasets derived from animal data.

| Dataset | Protocol Name | Location | Description |
|-----------|----------------------------------|---|--|
| Crowd | In-Clinic Crowd-Sourcing | Banfield Pet Hospital clinics | Dogs that were awaiting in-clinic care were outfitted with activity monitors and were video recorded for several hours each performing typical in-kennel behaviors. |
| | PI Participant Crowd-Sourcing | Multi-source: For instance, at-home, in-car, during walks and hikes. | Participants in the Pet Insight Project used smartphones to video record their dogs while wearing activity monitors in every-day situations. Collar fit, device orientation, environment, and animal behavior were meant to be representative of real-world usage. |
| Eat/drink | Waltham Eat/Drink Study | WALTHAM Petcare Science Institute | Dogs were video recorded eating and drinking by researchers at the WALTHAM Petcare Science Institute. Collar fit and orientation were controlled, and dog behaviors exhibited were relatively consistent. |

For brevity, we refer to these datasets simply as the *crowd* and *eat/drink* datasets.

2.4. Eat/Drink Study Protocol

This study was conducted using dogs owned by the WALTHAM Petcare Science Institute and housed in accordance with conditions stipulated under the UK Animals (Scientific Procedures) Act 1986. Briefly, the dogs were pair housed in environmentally enriched kennels designed to provide dogs free access to a temperature-controlled interior and an external pen at ambient temperature. Dogs were provided with sleeping platforms at night. The dogs had access to environmentally enriched paddocks for group socialization and received lead walks and off-lead exercise opportunities during the day. Water was freely available at all times and dogs were fed to maintain an ideal body condition score. The study was approved by the WALTHAM Animal Welfare and Ethical Review Body. One hundred and thirty-eight dogs across 5 different breeds (72 Labrador Retrievers, 18 Beagles, 17 Petit Basset Griffon Vendeens, 14 Norfolk Terriers and 17 Yorkshire Terriers) took part for two consecutive days each. Each dog was recorded once a day during its normal eating and drinking routine using a GoPro camera (GoPro, San Mateo, CA, USA).

In this study, either one (ventral only) or four (ventral, dorsal, left, and right) activity monitors were affixed to a collar. For each observation, the collar was removed from the dog, the correct number of activity monitors were attached, and then shaken sharply in view of the camera to provide a synchronization point that was identifiable in both the video and accelerometer signals (so that any time offset could be removed). The collar was then placed on the dog at a standardized tightness. The dogs were recorded from approximately one minute before feeding until approximately one minute after feeding. In order to increase the diversity of the dataset, collar tightness was varied between a two-finger gap and a four-finger gap, and food bowls were rotated between normal bowls and slow-feeder or puzzle-feeder bowls. For each data recording, researchers noted the date and time, device serial number(s), collar tightness, food amount and type, and various dog demographic data.

2.5. Crowd-Sourcing Protocol

Pet Insight participants were requested to use smartphones to video record their pets performing everyday activities while wearing activity monitors. The participants were told that the activity monitor should be worn on the collar but were not given any other instructions about how the collar or monitor should be worn. Participants were asked to prioritize recording health-related behaviors like scratching or vomiting, but to never induce these events and to never delay treatment in order to record the behaviors. As a participation incentive, for every crowd-sourced video used, the PI project donated one dollar to a pet-related charity.

After recording each video, participants logged into the PI crowd-sourcing website, provided informed consent, uploaded the recorded video, and completed a short questionnaire confirming which pet was recorded and whether certain behaviors were observed. The device automatically uploaded its accelerometry data to Whistle servers.

2.6. In-Clinic Observational Protocol

This study was conducted at several Banfield Pet Hospital (BPH) clinics. Its objective was to acquire long-duration (multi-hour) naturalistic recordings to augment the shorter crowd-sourced recordings, which were typically several minutes or less in duration.

Randomly selected BPH clients who chose to participate signed fully informed consent forms. Their dogs were outfitted with Velcro breakaway collars with one attached activity monitor device each. Collar tightness and orientation were not carefully controlled. Video was recorded via a 4-channel closed-circuit 720 p digital video security system. Video cameras were ceiling- or wall-mounted and oriented towards the in-clinic kennels so that up to four dogs could be observed at a time. For each recording, researchers noted the date and time, the device serial number, and the dog/patient ID number.

2.7. Video Labeling

All uploaded videos were transcoded into a common format (H.264-encoded, 720 p resolution, and up to 1.6 Mb/s) using Amazon's managed Elastic Transcoder service, and their audio was stripped for privacy. Video start times were extracted from the video metadata and video filenames. Matching device accelerometry data were downloaded from Whistle's databases, and automatic quality checks were performed.

Videos were then labeled by trained contractors using the open-source BORIS (Behavioral Observation Research Initiative Software V. 7.9.8) software application [33]. The resulting event labels were imported and quality-checked using custom Python scripts running on one of the PI project's cloud-based web servers. Labels were stored alongside video and participant metadata in a PostgreSQL database.

All video labeling contractors were trained using a standardized training protocol, and inter-rater reliability analyses were performed during training to ensure consistent labeling. Videos were labeled according to a project ethogram [8,15,20]. This report describes several of these label categories.

Labelers divided each video into *valid* and *invalid* regions. Regions were only valid if the dog was clearly wearing an activity monitor and was fully and clearly visible in the video. Invalid regions were subsequently ignored. In each *valid* video region, the labeler recorded exactly one *posture*, and any number (0 or more) of applicable *behaviors*.

Postures (Table 2) reflect the approximate position and energy expenditure level of the pet, while *behaviors* (Table 3) characterize the pet's dominant behavior or activity in a given moment. For instance, during a meal, a dog might exhibit a STAND posture and an EAT behavior. While pausing afterwards, the same dog might exhibit a STAND posture and no behavior. Multiple simultaneous behaviors are rare but possible, such as simultaneous SCRATCH and SHAKE behaviors.

Table 2. Postures Ethogram.

| Posture | Activity |
|----------|--|
| LIE DOWN | Lying down. |
| SIT | Sitting still with little to no movement |
| STAND | Standing still with little to no movement. |
| WALK | Purposeful walking from one point to another. |
| VIGOROUS | Catch-all for high-energy activities such as running, swimming, and playing. |
| MIXED | Default category for any other posture, for ambiguous postures, and for postures that are difficult to label due to rapid changes. |

Table 3. Behaviors Ethogram.

| Behavior | Activity |
|------------|---|
| DRINK | Drinking water. |
| EAT | Eating food, as in out of a bowl. Does not include chewing bones or toys. |
| LICKOBJECT | Licking an object other than self, such as a person or empty bowl. |
| LICKSELF | Self-licking, often due to pain, soreness, pruritis, or trying to clear a foreign object. |
| PETTING | Being pet by a human. |
| RUBBING | Rubbing face or body on an object or person due to pruritis. |
| SCRATCH | Scratching of the body, neck, or head with a hind leg. |
| SHAKE | Shaking head and body, as in when wet. Does not include head-shaking that is clearly due to ear discomfort, which is labeled separately and has not been included in this report. |
| SNIFF | Sniffing the ground, the air, a person or other pet |
| NONE | 'Default' class indicating that no labeled behavior is happening. |

2.8. Training Data Preparation

Although accelerometer data and smartphone video data were both time-stamped using the devices' network-connected clocks, inaccuracies led to alignment errors of typically several seconds, and sometimes much longer. Short activities such as SHAKE, in particular, require more accurate alignment. We aligned approximately 1200 videos manually by matching peaks in accelerometer activity to labels for high-intensity behaviors like SHAKE and SCRATCH. We used these manual alignments to develop and validate an automatic alignment algorithm that aligned the remaining videos.

We created each of the two training datasets (*crowd* and *eat/drink*) by:

1. Selecting appropriate videos from our database.
2. Limiting the number of entries per dog to 30 (some dogs are overrepresented in our database).
3. Allocating all of each dog's data into one of 5 disjoint cross-validation folds.
4. Downloading each dataset and labeling each time-point with a posture and/or behavior(s).

The specific method of separating data into cross-validation folds (step 3 above) is critical [34]. Classifiers trained on individual dogs have been shown to over-perform on those dogs relative to others, even if those classifiers are trained and evaluated using separate experimental observations. Gerencsér et al. experienced an accuracy reduction from 91% for a single-subject classifier to 70–74% when generalizing to other dogs [35]. Consequently, we were careful to ensure that all of a dog's videos fall in a single fold, so that data from a single dog is never used to both train and evaluate a classifier.

The overall data acquisition process, from video capture (red), accelerometer data (blue) to a completed dataset (purple), is shown in Figure 2.

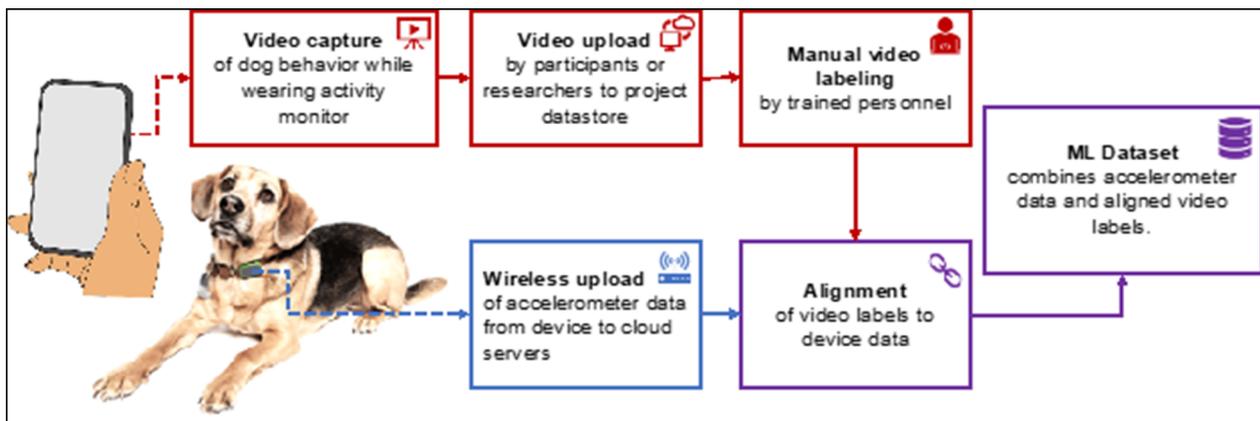


Figure 2. Data acquisition flow. Dogs wearing collar-mounted activity monitors were video recorded performing behaviors of interest or performing everyday activities. Videos were uploaded and the behaviors exhibited in them were manually labeled (tagged). The devices automatically uploaded accelerometer (activity) data to cloud servers, and the device data were aligned with the video labels to remove any temporal offset. The aligned labels and accelerometer time series were combined into datasets suitable for training machine learning (ML) models.

2.9. Deep Learning Classifier

Our deep learning classifier is based on our FilterNet architecture [32]. We implemented the model in Python using PyTorch v1.0.1 [36] and the 2020.02 release of the Anaconda Python distribution (64-bit, Python 3.7.5). We trained and evaluated our models on p2.xlarge instances on Amazon Web Services [37] with 4 vCPUs (Intel Xeon E5-2686 v4), 61 GB RAM, and a NVIDIA Tesla k80 GPU with 12 Gb RAM, running Ubuntu 18.04.4.

We used the *crowd* dataset for cross-validated training and evaluation (Figure 3). Specifically, we trained and evaluated five different models, using a different held-out fold as a test set for each model. We combine the models' predictions for each of the five test sets for model evaluation, as described below. We also generated behavior classifications for the *eat/drink* dataset using one of the models trained on the *crowd* dataset (that is, we did not use the *eat/drink* dataset for model training). There were no dogs in common between the *crowd* and *eat/drink* datasets, so cross-validation was not needed in this step.

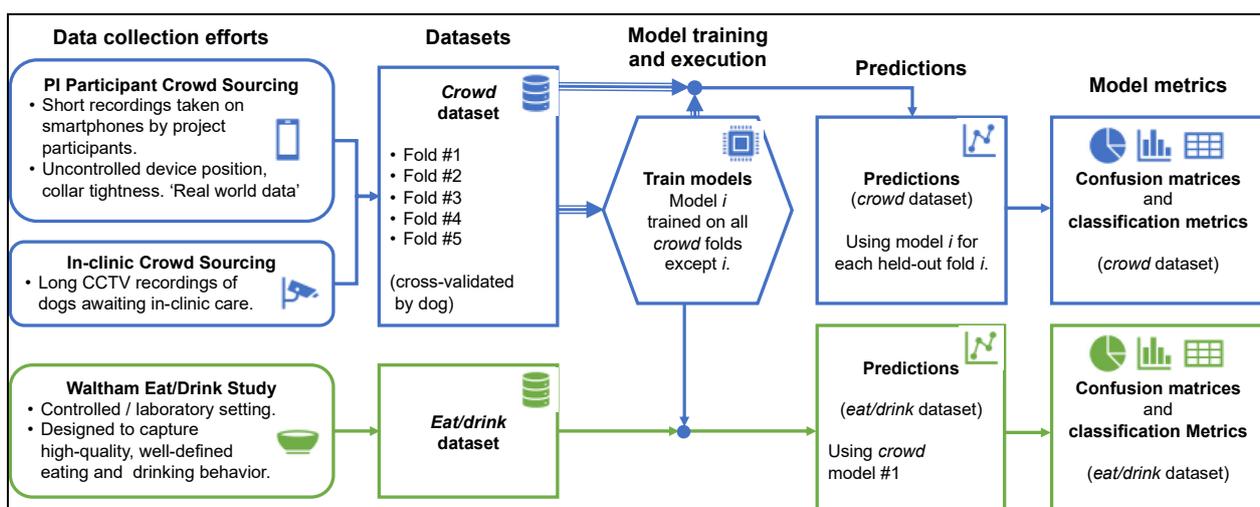


Figure 3. Model training and evaluation data flow. The *crowd* dataset consisted of naturalistic, highly diverse data divided by dog into five folds. The *eat/drink* dataset focused on high-quality eating and drinking data. Behavior classification models were trained and evaluated in a cross-validated fashion (where a given model i is trained on all folds of data except fold i) on the *crowd* dataset, and the first of these five models was also evaluated on the *eat/drink* dataset. Confusion matrices and classification metrics were produced for each dataset using the resulting predictions.

2.10. Evaluation

For evaluation, we modeled the task as two multi-class classification problems, one for behaviors and one for postures. At each time point in each video entry in a dataset we recorded the *labeled* behavior and posture, and every 320 ms we calculated the most likely *predicted* behavior and posture. We tallied the labeled and predicted pairs from all five test folds together using the *PyCM* multiclass confusion matrix library to create separate behavior and posture confusion matrices [38]. We used the *PyCM* package to calculate metrics derived from the confusion matrices [39].

As the *MIXED* posture is used primarily for expediency in labeling, we dropped any time points with *MIXED* labels from the postures confusion matrix, and replaced any *MIXED*-class posture predictions with the next most likely prediction for that time point. We also excluded any time points with more than one simultaneous labeled behavior (about 3% of the data) from the behaviors confusion matrix.

Furthermore, following Uijl et al. [8], we excluded any time points within 1 s of a class transition in both classification problems. However, also similar to [8], we treated the *SHAKE* class differently due to its very short duration. For *SHAKE*, we only excluded the outer one-third second. In dropping these transition regions, we attempted to follow established convention for minimizing the effects of misalignment in labeling, and to make our reported results easier to compare to related works.

Performance of these models was evaluated based on widely used metrics in the machine learning field including F1 scores. These metrics can be expressed in terms of the number of true and false positive predictions (*TP* and *FP*) and the number of true and false negative predictions (*TN* and *FN*). They include precision ($TP/(TP + FP)$), sensitivity or recall ($TP/(TP + FN)$), and specificity ($TN/(TN + FP)$). F1 scores examine the relationship between the precision and recall of a model to better understand a model's accuracy.

2.11. User Validation

Although the *crowd* dataset is meant to be representative of real-world data, it is subject to biases such as underrepresentation of behaviors that are unlikely to be video recorded, such as riding in cars or staying at home alone. Furthermore, it is impossible to anticipate all of the myriad situations that may serve as confounders. Consequently, we ran real-world user validation campaigns on the two behaviors that users are most likely to be aware of, *EAT* and *DRINK* behavior. We defined events as periods of relatively sustained, specific behaviors detected with high confidence, such as eating events (meals) consisting of several minutes of sustained eating behavior. We adapted our production system, which runs the models described in this study in near-real time on all PI project participants, to occasionally send validation emails to participants when an *EAT* or *DRINK* event had occurred within the past 15 min. Respondents categorized the event detection as correct ("Yes") or incorrect ("No") or indicated that they were not sure. Users were able to suggest what confounding event may have triggered any false predictions. We excluded any responses that arrived more than 60 min after an event's end, as well as any "Not Sure" responses.

3. Results

3.1. Data Collected

After applying the steps described above, the *crowd* dataset contained data from 5063 videos representing 2217 subjects, and the *eat/drink* dataset contained data from 262 videos representing 149 unique dogs. The distribution of weights and ages represented in these datasets is shown in Figure 4, while a breed breakdown is given in Table 4.

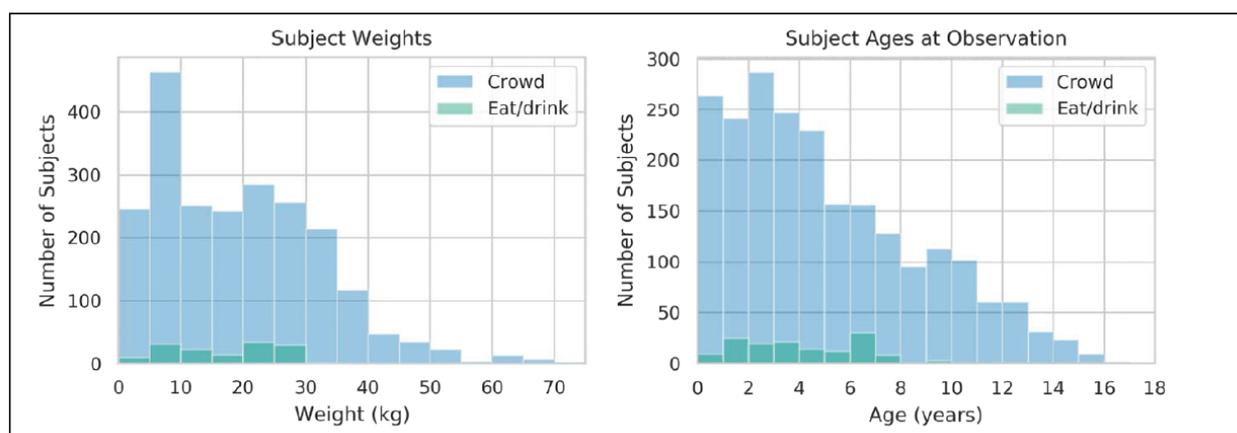


Figure 4. Weight and age distribution for the *crowd* and *eat/drink* datasets. Due to its more heterogeneous population, the *crowd* dataset had much greater variability than the *eat/drink* dataset.

Table 4. Breed or breed category breakdown for the *crowd* and *eat/drink* datasets.

| Breed | Crowd | Eat/Drink |
|------------------------------|-------|-----------|
| Mixed | 923 | 0 |
| Unknown/Other | 413 | 0 |
| Beagle | 38 | 18 |
| Boxer | 26 | 0 |
| Bulldog | 44 | 0 |
| Collie | 23 | 0 |
| Great Dane | 25 | 0 |
| Hunting dog | 43 | 0 |
| Norfolk Terrier | 0 | 14 |
| Petit Basset Griffon Vendéen | 0 | 17 |
| Pug | 39 | 0 |
| Retriever | 128 | 72 |
| Shepherd | 68 | 0 |
| Sled dog | 44 | 0 |
| Small dog | 189 | 0 |
| Spaniel | 20 | 0 |
| Yorkshire Terrier | 195 | 17 |
| # Unique breeds | 338 | 0 |

These datasets also differed in the length and frequency of labeled events, as shown in Table 5. The *crowd* and *eat/drink* datasets contain 163.9 and 22.4 h of video data labeled as VALID, respectively.

The EAT class was highly represented in both the *crowd* dataset (because participants were specifically requested to submit videos of their dogs at mealtime, since it is an easily filmed and important behavior) and in the *eat/drink* dataset (due to study design). The *eat/drink* dataset included only small amounts of incidental LICKSELF, SCRATCH, PETTING, and SHAKE behavior, while the *crowd* dataset contained many of these events because participants were repeatedly reminded of their importance.

The distribution of lengths for each label class was highly skewed, with many short labels and a smaller number of longer labels (Figure 5). The distribution of SHAKE labels was less skewed, likely because it is typically a short behavior and less prone to interruption.

Table 5. Summary of labeled crowd and eat/drink datasets.

| No. Valid Videos | Eat/Drink Dataset | | | | Crowd Dataset | | | |
|------------------|-------------------|----------|-----------------|----------------------|---------------|----------|-----------------|----------------------|
| | # Videos | # Labels | Mean Length (s) | Total Length (H:M:S) | # Videos | # Labels | Mean Length (s) | Total Length (H:M:S) |
| | 5367 | 10,529 | 56 | 163:51:46 | 262 | 409 | 188 | 21:26:15 |
| Behaviors | | | | | | | | |
| DRINK | 1072 | 1586 | 12.6 | 5:33:14 | 151 | 311 | 5 | 0:25:52 |
| EAT | 1484 | 3063 | 43.4 | 36:57:31 | 260 | 375 | 65 | 6:46:33 |
| LICKOBJECT | 897 | 2056 | 5.6 | 3:11:45 | 121 | 212 | 5.7 | 0:20:18 |
| LICKSELF | 464 | 1405 | 12.4 | 4:50:11 | 8 | 8 | 7.3 | 0:00:58 |
| SCRATCH | 416 | 679 | 6.8 | 1:16:41 | 4 | 5 | 4.0 | 0:00:20 |
| PETTING | 510 | 901 | 6.7 | 1:41:03 | 72 | 92 | 2.0 | 0:04:04 |
| RUBBING | 271 | 605 | 7.0 | 1:10:32 | 0 | 0 | 0 | 0:00:00 |
| SHAKE | 552 | 689 | 1.7 | 0:19:09 | 48 | 64 | 1.0 | 0:1:04 |
| SNIFF | 2668 | 9696 | 3.9 | 10:31:56 | 241 | 1082 | 5.0 | 1:36:46 |
| Postures | | | | | | | | |
| LIE DOWN | 1340 | 2407 | 65.7 | 43:56:24 | 10 | 11 | 58.8 | 0:10:46 |
| MIXED | 4072 | 11,538 | 9.9 | 31:49:27 | 261 | 1492 | 30.0 | 12:44:22 |
| VIGOROUS | 633 | 1968 | 8.7 | 4:45:24 | 15 | 24 | 6.9 | 0:02:44 |
| SIT | 1500 | 2784 | 19.8 | 15:18:20 | 136 | 263 | 21.0 | 1:32:42 |
| STAND | 3725 | 9090 | 22.5 | 56:44:37 | 234 | 1294 | 19.0 | 6:55:51 |
| WALK | 1255 | 4101 | 9.9 | 11:16:34 | 5 | 21 | 6.0 | 0:02:06 |

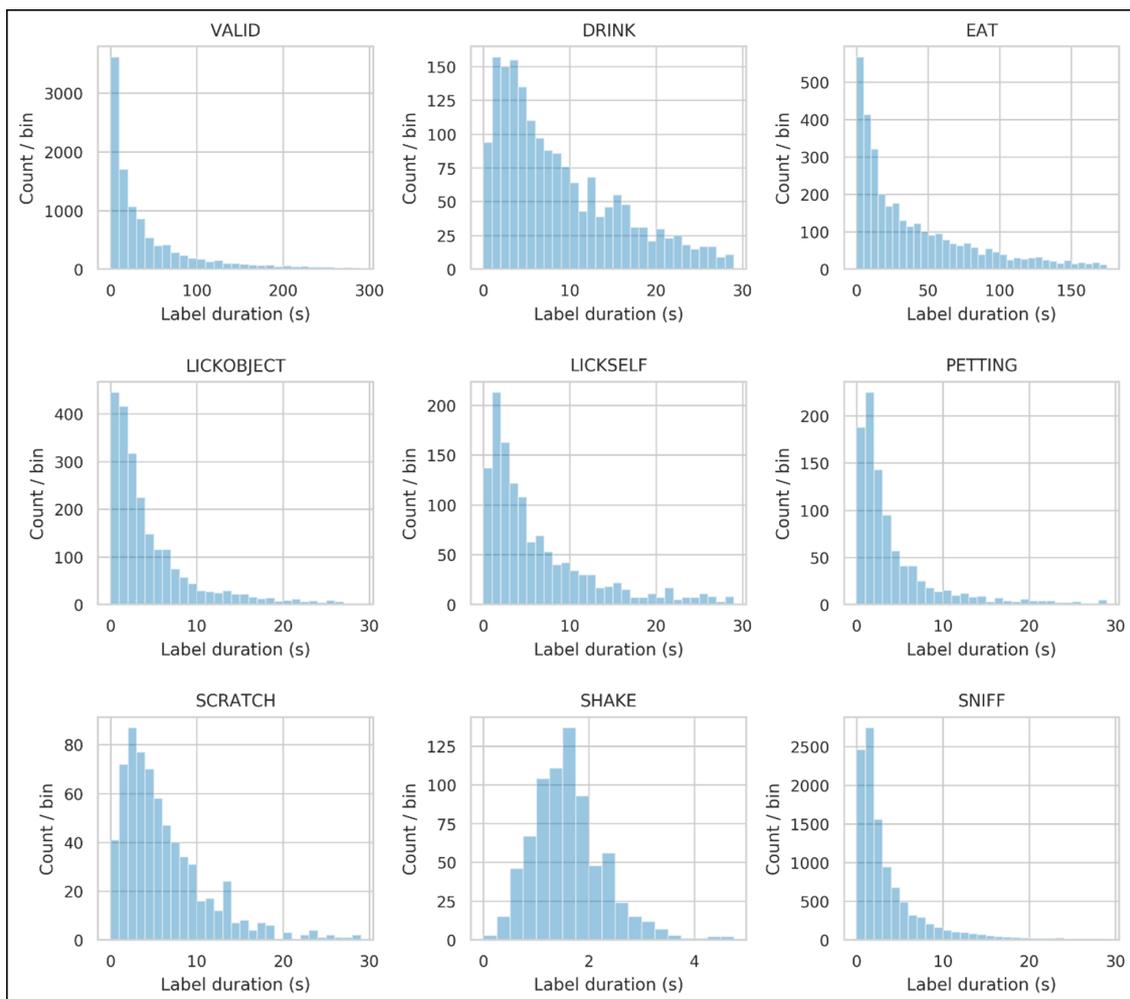


Figure 5. Distribution of label durations in the combined eat/drink and crowd datasets. Except for SHAKE, all labels exhibit a highly skewed distribution, probably the result of some longer label segments becoming fragmented.

3.2. Classification Accuracy

Cross-validated classification metrics for the *crowd* dataset are given in Table 6, and classification metrics obtained from evaluating the *eat/drink* dataset using a model trained on the *crowd* dataset are given in Table 7. Subsequent sections may report behaviors only due to postures having less accurate labels and are typically used in an aggregate form where individual misclassifications are less important.

Table 6. Classification metrics for the crowd dataset.

| Behavior | # Dogs | # Videos | Prevalence (Support) | Sensitivity (Recall) | Specificity | Accuracy | Precision (PPV) | F1 Score |
|-----------------|--------|----------|----------------------|----------------------|-------------|----------|-----------------|----------|
| <i>Behavior</i> | | | | | | | | |
| DRINK | 752 | 1019 | 3.9% | 0.874 | 0.995 | 0.99 | 0.870 | 0.872 |
| EAT | 1101 | 1442 | 28% | 0.902 | 0.967 | 0.948 | 0.915 | 0.908 |
| LICKOBJECT | 460 | 629 | 1.8% | 0.410 | 0.990 | 0.98 | 0.439 | 0.424 |
| LICKSELF | 257 | 398 | 3.4% | 0.772 | 0.990 | 0.982 | 0.728 | 0.749 |
| PETTING | 204 | 307 | 0.96% | 0.305 | 0.991 | 0.984 | 0.237 | 0.267 |
| RUBBING | 158 | 235 | 0.73% | 0.729 | 0.996 | 0.994 | 0.584 | 0.648 |
| SCRATCH | 158 | 303 | 0.60% | 0.870 | 0.997 | 0.997 | 0.676 | 0.761 |
| SHAKE | 251 | 435 | 0.13% | 0.916 | 1.00 | 1.00 | 0.795 | 0.851 |
| SMELL | 946 | 1747 | 5.3% | 0.610 | 0.968 | 0.949 | 0.517 | 0.559 |
| NONE | 2051 | 4636 | 55% | 0.892 | 0.898 | 0.895 | 0.914 | 0.903 |
| <i>Posture</i> | | | | | | | | |
| LIE DOWN | 674 | 1223 | 22% | 0.826 | 0.913 | 0.894 | 0.724 | 0.772 |
| SIT | 726 | 1275 | 9.9% | 0.409 | 0.915 | 0.865 | 0.347 | 0.375 |
| STAND | 2028 | 4241 | 58% | 0.793 | 0.900 | 0.838 | 0.916 | 0.850 |
| VIGOROUS | 289 | 468 | 2.9% | 0.764 | 0.985 | 0.978 | 0.605 | 0.675 |
| WALK | 599 | 905 | 7.7% | 0.903 | 0.969 | 0.964 | 0.706 | 0.793 |

Table 7. Classification metrics for the *eat/drink* dataset *.

| | # Dogs | # Videos | Prevalence (Support) | Sensitivity (Recall) | Specificity | Accuracy | Precision (PPV) | F1 Score |
|-----------------|--------|----------|----------------------|----------------------|-------------|----------|-----------------|----------|
| <i>Behavior</i> | | | | | | | | |
| DRINK | 71 | 99 | 1.7% | 0.949 | 0.999 | 0.998 | 0.957 | 0.953 |
| EAT | 147 | 259 | 33% | 0.988 | 0.983 | 0.984 | 0.966 | 0.977 |
| LICKOBJECT | 70 | 95 | 1.6% | 0.658 | 0.998 | 0.992 | 0.821 | 0.731 |
| SMELL | 142 | 231 | 5% | 0.780 | 0.981 | 0.971 | 0.681 | 0.728 |
| NONE | 149 | 262 | 58% | 0.959 | 0.968 | 0.963 | 0.977 | 0.968 |
| <i>Posture</i> | | | | | | | | |
| SIT | 79 | 79 | 11% | 0.447 | 0.940 | 0.886 | 0.481 | 0.464 |
| STAND | 149 | 262 | 88% | 0.938 | 0.469 | 0.883 | 0.930 | 0.934 |

* Not reported for categories with <0.05% support, <25 dogs, or <50 videos.

Of the metrics in Tables 6 and 7, only sensitivity and specificity are independent of class prevalence.

The “behaviors” confusion matrix for the *crowd* dataset is shown in Figure 6 in non-normalized and normalized forms. The non-normalized confusion matrix gives raw tallies (that is, the total number of one-third second time points) of predicted and labeled classes, and the normalized confusion matrix gives the percentage of each actual label classified by the algorithms as a given predicted label (so that the percentages in each row sum to 100%). The non-normalized matrix is dominated by correctly predicted NONE and EAT samples, due to their high prevalence and effective classification in this dataset. The normalized matrix suggests the reliable classification of DRINK, EAT, NONE, and SHAKE. The LICKSELF and SCRATCH classes are of moderate reliability, and the LICKOBJECT,

PETTING, RUBBING, and SNIFF classes exhibit some systematic misclassification and are of lesser reliability.

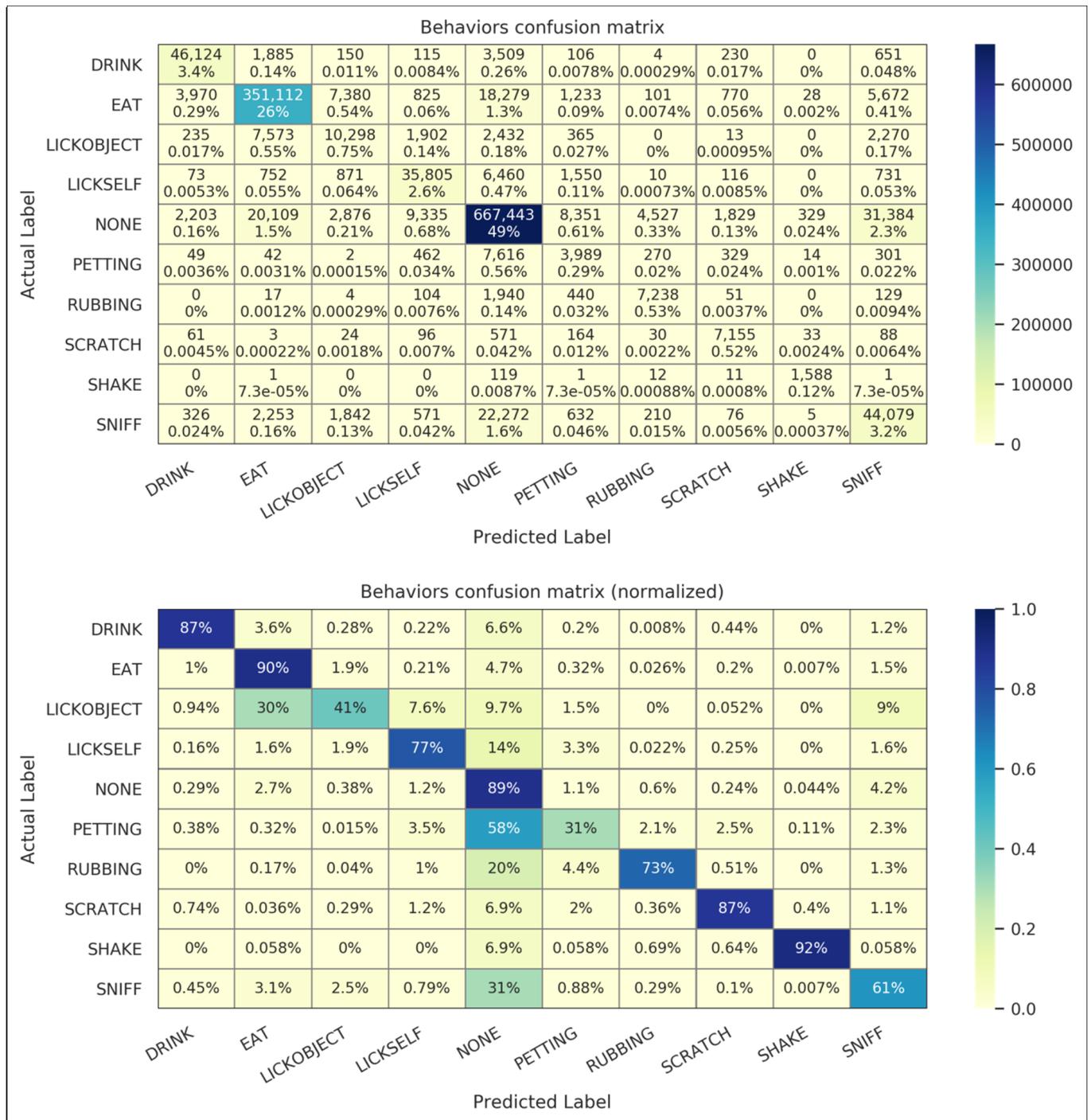


Figure 6. Confusion matrices for crowd dataset behaviors. Time point values in raw counts (top) and normalized by time point labels (bottom) are shown.

3.3. Effect of Device Position on Performance

The system’s classification performance, as measured by F1 score, shows no significant dependence on device position (Figure 7).

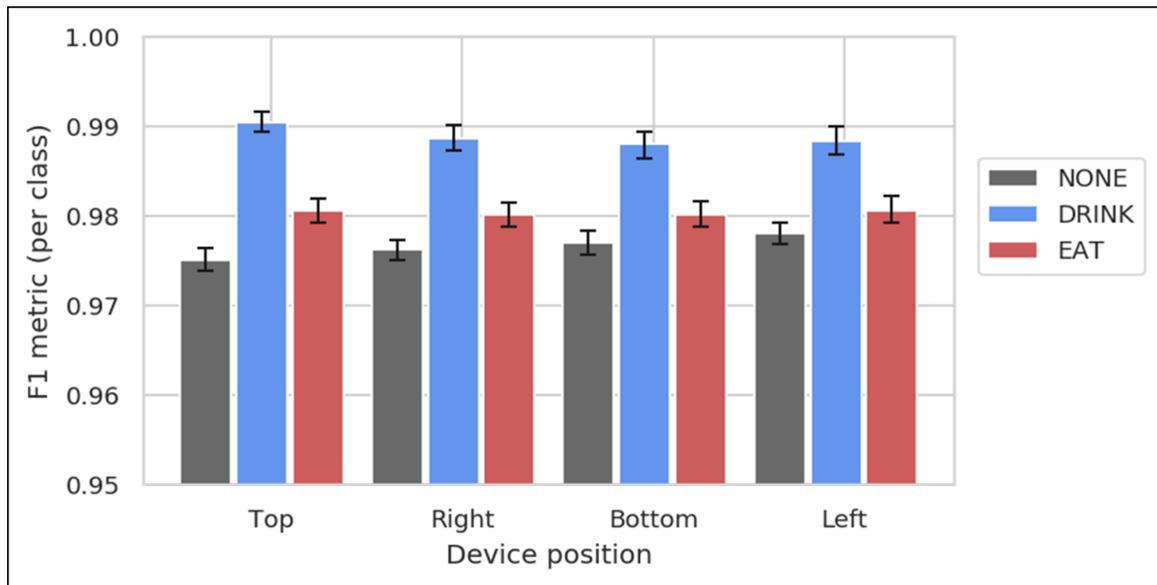


Figure 7. Classification performance as measured by F1 score. F1 scores measuring test accuracy are broken out by device position, for $n = 48$ videos from the *eat/drink* dataset where the dog’s collar had exactly four attached devices with known orientation. Error bars are 95% confidence intervals on the mean, as determined by bootstrapping. The classification accuracy per class was similar between the four positions, indicating that system accuracy is not substantially affected by collar rotation.

3.4. User Validation

Participants responded far better than expected to user validation efforts. Users opened emails, clicked through to the web form, and submitted validation results for 55% of the EAT validation emails and 42% of the DRINK validation emails.

Responses are summarized in Table 8. As described above, we excluded any responses that arrived more than 60 min after an event’s end, as well as any “Not Sure” responses. The positive (“Yes”) validation rate was approximately 95% for both event types. As expected, the rate of users responding “Not Sure” was far greater for DRINK (12%) than for EAT (2%).

Table 8. Summary of user EAT and DRINK validation responses.

| Event | Users | Valid Responses | | | Confounders |
|-------|-------|-----------------|------|-------|---|
| | | Response | N | % | |
| EAT | 1514 | Yes | 2488 | 95.3% | Drinking, vomiting/regurgitation, eating grass, licking (pan/bowl/self), chewing (bone/toy), playing. |
| | | No | 123 | 4.7% | |
| DRINK | 1491 | Yes | 2579 | 94.9% | Eating, licking, sniffing, chewing (bone/toy), petting. |
| | | No | 140 | 5.1% | |

As the production system generates candidate EAT and DRINK events, it calculates a confidence score (the mean algorithm confidence over the event’s duration) that varies between 0 and 1.0, and drops any events with a score below a threshold of 0.3. Figure 8 shows how the percentage of “Yes” responses (the true positive rate) varied with this confidence score. For EAT events, the rate grew from 83% for the lowest-confidence bin (0.3–0.4) to 100% (201 out of 201) for the highest-confidence bin (0.9–1.0). Since users do not see the confidence score, this trend suggests that the EAT validation data are relatively reliable. The DRINK data show a less convincing trend, which is consistent with users’ lower awareness of DRINK events.

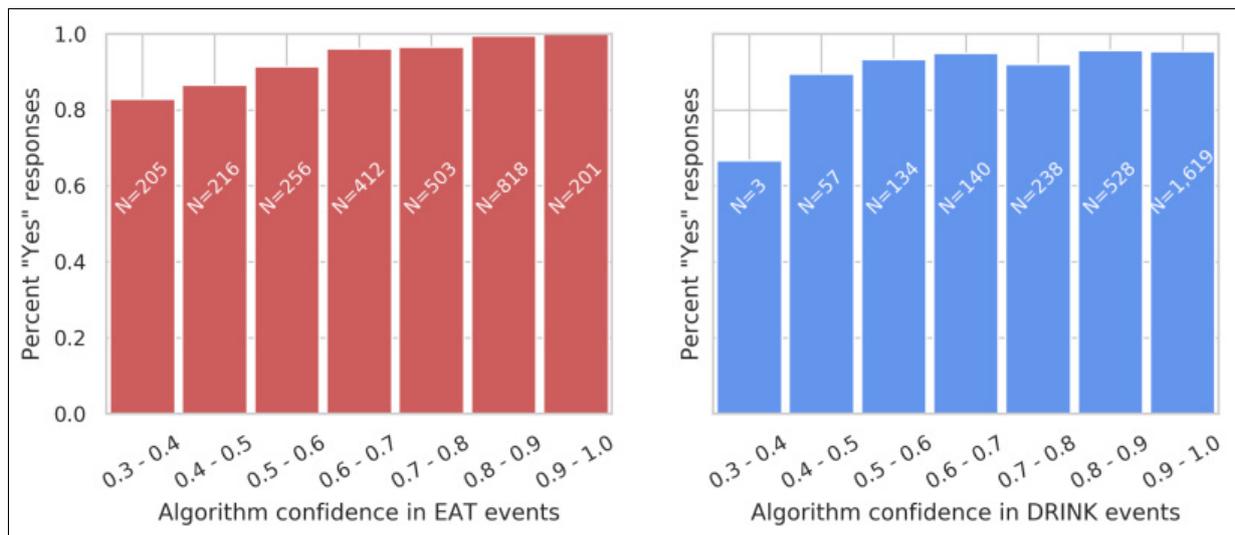


Figure 8. User responses to EAT and DRINK event validation requests. Responses to EAT events (**left**) and DRINK events (**right**) are grouped by the algorithm's confidence in each event. Validation requests were sent automatically by email within 15 min of an event, and responses were allowed within 1 h of an event. Confidence estimates varied from 0.3 to 1.0 (events with lower confidence are dropped by the algorithm). Bars are annotated with the total number of responses in each bin.

It is unfortunate that, of the behavior classes measured in this work, only EAT is likely to exhibit the level of user awareness required for validation using this method.

4. Discussion

4.1. Comparison with Previous Work

We compared our dataset and results with several previous works (Table 9), and we tabulated several important qualitative differences between the datasets (Table 10). In comparing these results, it is important to account for:

- *Class distribution.* Each dataset exhibits a different distribution of behaviors. In general, classifiers exhibit better F1 scores for common behaviors than for rare behaviors. The classifier sensitivity and specificity are relatively insensitive to this distribution, so we recommend using these metrics for comparing performance across different datasets.
- *Dataset collection methods.* Classifiers are more accurate when applied to high-quality datasets collected under controlled conditions. Accuracy can drop substantially in naturalistic versus laboratory settings [26,27]. Classifiers benefit from consistent device position, device attachment, and collar tightness, and they also benefit when the labeled behaviors as well as the collection environment are consistent and well-defined.

Previous works have used relatively controlled and high-quality datasets, similar to the *eat/drink* dataset in this work [8,18,19,21]. As expected, our crowd sourced dataset exhibits a far greater diversity of weights, ages, and breeds than our *eat/drink* dataset, since the *eat/drink* subjects are sampled from several relatively homogeneous subpopulations.

The classification performance of the classifier presented here on the EAT and DRINK classes in the *eat/drink* dataset advances the sensitivity, specificity, and F1 score for these classes. Sensitivity and specificity are independent of class prevalence. The balance between sensitivity and specificity is a design choice, so we have calibrated our algorithms to favor specificity in order to minimize false positives.

Table 9. Comparison of this work to other published results.

| Reference | | # Dogs | Prevalence (Support) | Sensitivity (Recall) ¹ | Specificity ¹ | Accuracy | Precision (PPV) | F1 Score |
|---------------------------------------|------------|--------|----------------------|-----------------------------------|--------------------------|----------|-----------------|----------|
| <i>crowd</i> dataset (this study) | DRINK | 752 | 3.9% | 0.874 | 0.995 | 0.990 | 0.870 | 0.872 |
| | EAT | 1101 | 28% | 0.902 | 0.967 | 0.948 | 0.915 | 0.908 |
| | LICKOBJECT | 460 | 1.8% | 0.410 | 0.990 | 0.980 | 0.439 | 0.424 |
| | LICKSELF | 257 | 3.4% | 0.772 | 0.990 | 0.982 | 0.728 | 0.749 |
| | PETTING | 204 | 0.96% | 0.305 | 0.991 | 0.984 | 0.237 | 0.267 |
| | RUBBING | 158 | 0.73% | 0.729 | 0.996 | 0.994 | 0.584 | 0.648 |
| | SCRATCH | 158 | 0.60% | 0.870 | 0.997 | 0.997 | 0.676 | 0.761 |
| | SHAKE | 251 | 0.13% | 0.916 | 1.000 | 1.000 | 0.795 | 0.851 |
| <i>eat/drink</i> dataset (this study) | SNIFF | 946 | 5.3% | 0.610 | 0.968 | 0.949 | 0.517 | 0.559 |
| | DRINK | 71 | 1.7% | 0.949 | 0.999 | 0.998 | 0.957 | 0.953 |
| | EAT | 147 | 33% | 0.988 | 0.983 | 0.984 | 0.966 | 0.977 |
| | LICKOBJECT | 70 | 1.6% | 0.658 | 0.998 | 0.992 | 0.821 | 0.731 |
| Griffies et al. 2018 [19] | SNIFF | 142 | 5% | 0.780 | 0.981 | 0.971 | 0.681 | 0.728 |
| | SCRATCH | | 2.12% | 0.769 | 0.997 | 0.992 | 0.861 | 0.812 |
| den Uijl et al. 2016 [18] | SHAKE | | 0.8% | 0.722 | 0.998 | 0.996 | 0.726 | 0.724 |
| | DRINK | 206 | | 0.76 | 0.97 | | 0.86 | 0.81 |
| | EAT | 242 | | 0.77 | 0.97 | | 0.84 | 0.80 |
| den Uijl et al. 2017 [8] | SHAKE | 145 | | 0.91 | 1.00 | | 0.79 | 0.85 |
| | DRINK | 23 | | 0.89 | 0.87 | 1.00 | 1.00 | 0.94 |
| | EAT | 23 | | 0.92 | 0.73 | 0.99 | 0.73 | 0.81 |
| Kiyohara et al. 2015 [21] | SHAKE * | 51 | | 0.98 | 0.95 | 0.95 | 0.99 | 0.98 |
| | DRINK | 2 | | 0.02 | | | 0.14 | 0.04 |
| | EAT | 2 | | 0.28 | | | 0.34 | 0.31 |

¹ Best sensitivity and specificity for each class are in boldface. * SHAKE registered *per-event* as opposed to *per-time-sample* in this analysis.

Table 10. Dataset collection methods in similar published studies.

| Reference * | Device Position | Collar Fit | Environment | Behaviors |
|---------------------------------------|-----------------|--------------|--------------------------------------|-----------------|
| <i>Crowd</i> dataset (this study) | Uncontrolled | Uncontrolled | Uncontrolled (highly varied) | Naturalistic |
| <i>Eat/drink</i> dataset (this study) | Controlled | Controlled | Controlled (lab/kennel) | Semi-controlled |
| Griffies et al., 2018 [19] | Controlled | Controlled | Controlled (animal shelter) | Naturalistic |
| den Uijl et al., 2017 [8] | Controlled | Controlled | Controlled (track/field and indoors) | Semi-controlled |

* Relevant data collection details not available for den Uijl et al. 2016 [18] or Kiyohara et al. 2015 [21].

The classifiers' performance on SCRATCH in the challenging *crowd* dataset also advances the state of the art. Comparable detection of LICKOBJECT, LICKSELF, PETTING, RUBBING, and SNIFF has not been previously demonstrated to our knowledge. We note that SCRATCH, LICKSELF, and RUBBING behaviors are highly relevant to dermatological health and welfare applications [19], and that PETTING is an important confounder that can be easily misclassified as SCRATCH or LICKSELF in classifiers that are not exposed to this behavior. We have found the classifiers' detection of SHAKE to be highly accurate (though susceptible to temporal misalignment between device and video data, due to the short event lengths). It is difficult to compare the *per-time-sample* SHAKE classification metrics here to published *per-event* metrics due to differing methodologies [8,18].

The device position invariance demonstrated by our classifier is a key property that enables real-world performance to approach that of controlled studies, allowing accurate detection of our reported behaviors in home environments.

4.2. Challenges

In Supplementary Materials, we include seven videos (Videos S1–S7) annotated with behavior classification predictions, as well as an explanatory figure (Figure S1) and table (Table S1), in order to demonstrate the system's operation. The system excels at certain clearly defined and easily recognizable activities, especially those repeating and universal movement patterns such as drinking (lapping), walking, running, shaking, and most eating behaviors. It also performs well on well-defined instances of scratching and self-licking.

Device positioning and collar tightness do not appear to have a strong effect on system accuracy, meaning that accurate behavior metrics can be acquired via normal activity monitor usage. An important feature of the devices described in this study is their insensitivity (invariance) to collar orientation or position (Figure 7). In real-world settings, and especially with lightweight devices such as the Whistle FIT, the device can be, and often is, rotated away from the conventional ventral (bottom) position at the lowest point of the collar.

The system appears to use the angle of a dog's neck (that is, whether the dog is looking up or down) as an important behavioral clue. Consequently, activities such as eating or drinking appear to be less accurate when raised dog bowls are used, and activities such as sniffing and scratching, and self-licking can go undetected if performed in unusual positions. Slow-feed food bowls, collars attached to taut leashes, and loose collars with other heavy attachments can also cause misclassifications, but are often classified correctly nonetheless.

The class distribution of both datasets is highly imbalanced, which presented a challenge for algorithm training. For instance, in the *crowd* dataset, which we used for training, the EAT class total duration is 117 times greater than that of SHAKE.

It is important to note that the class balance (class prevalence) of these datasets is not representative of real-world canine behavior. As the videos are typically taken in stimulating or interesting situations, these datasets exhibit a lower relative prevalence of LIE DOWN and other low-energy postures. Furthermore, the datasets exhibit much higher levels of EAT, DRINK, and possibly other behaviors, due to either study design (in the *eat/drink* dataset) or because the PI project requested that participants film certain behaviors.

Other sets of activities simply present very similar accelerometer data, such as eating wet food, which can be confounded with drinking; or being pet by a human or riding in a moving vehicle, which can be confounded with scratching or self-licking; or even vigorous playing and 'tug-of-war', which can be confounded with shaking and other activities. These misclassifications become less common as the models improve, but in some cases confusion may be unavoidable. Some other activities are simply rare or unusual, for instance, drinking from a stream, drinking from a water bottle, or licking food off of a raised plate.

A different type of problem relates to activities that are ambiguous even to human labelers, such as the distinction between eating a small part of a meal versus eating a large treat. Similarly, label fragmentation, where a long stretch of the labeled activity is interrupted either by the dog temporarily pausing (for instance, lifting up its head to look around several times while drinking or while eating a meal) or by discontinuities in the labeling when the dog leaves the camera's field of view (since labelers only marked videos as VALID when the dog was fully and clearly visible). These types of labeling ambiguity can be very deleterious to certain classification metrics, even though it is questionable whether the system's usefulness or real-world accuracy is affected.

User Validation participant comments confirmed our expectation that users were less aware of DRINK behavior than of EAT behavior. This lack of awareness likely also contributed to the lower DRINK response rate. It is unfortunate that, of the behavior classes measured in this work, only EAT is likely to exhibit the level of user awareness required for validation using this method.

5. Conclusions

We advanced the science of wearables through the development of novel machine learning algorithms which validated the sensitivity and specificity for detecting drinking and eating behavior. We also used a large real-world dataset of 2500 dogs to demonstrate detection of licking, petting, rubbing, scratching, and sniffing. Ensuring that the wearables would collect accurate data in a real-world setting, we demonstrated that system performance is not sensitive to collar position. In production, users reported high rates of true positives, consistent with the metrics measured via cross-validation on the *crowd* training database. This means that the data collected through the accelerometers in wearables can provide valuable data which can be applied in diagnosing and treating conditions. A subsequent survey of 10,550 dogs was used to validate the eating and drinking behavior. This survey takes the data from the laboratory and brings them into the real world to confirm results. The systems described in this work can further improve via the incorporation of additional training data and through the improvement of the underlying algorithms. Through the foundational algorithms built on the vast dataset, a world of opportunity is opened to further our understanding of animal behavior and advance individualized veterinarian care with the inclusion of wearables.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/ani11061549/s1>, Figure S1: Guide to interpreting supplemental videos, Table S1: Index of supplementary videos. Videos S1–S7 available online <https://zenodo.org/record/4836665#.YLR8s6ERVPY>.

Author Contributions: R.D.C., Conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft preparation; N.C.Y., conceptualization, data curation, formal analysis, investigation, methodology, software, validation, writing—review and editing; A.B.C., conceptualization, investigation—in-clinic crowd sourcing, writing—review and editing; C.J., data curation, investigation, methodology, software, writing—review and editing; D.E.A., conceptualization, data curation, investigation, validation, project administration; L.M.P., software, writing—review and editing; S.B., conceptualization support—eating and drinking study, investigation—experimental work at the WALTHAM Petcare Science Institute, writing—review and editing; G.W., conceptualization, funding acquisition, methodology, project administration; K.L., conceptualization, funding acquisition, methodology, resources; S.L., conceptualization, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: All funding for this work was provided by Mars Petcare.

Institutional Review Board Statement: This study was approved by the WALTHAM Animal Welfare and Ethical Review Body (Project Portfolio Management number 58565, June 2017) and conducted under the authority of the Animals (Scientific Procedures) Act 1986.

Data Availability Statement: Data available on request due to privacy restrictions. The data presented in this study are available on request from the corresponding author. The data is not publicly available.

Acknowledgments: We are grateful to Leonid Sudakov and Jeannine Taaffe for their support and vision in enabling the Pet Insight Project; WALTHAM Petcare Science Institute for contributing extensive training data and for numerous helpful discussions; to the many participants in the Pet Insight Project; and to the Whistle, Kinship, and Mars Petcare organizations for supporting the development and publication of this work.

Conflicts of Interest: All authors were employed by Mars Petcare during their contributions to this work.

References

1. Pewek, L.; Ellis, D.A.; Andrews, S.; Joinson, A. The rise of consumer health wearables: Promises and barriers. *PLoS Med.* **2016**, *13*, e1001953. [[CrossRef](#)]
2. Vogenberg, F.R.; Santilli, J. Healthcare trends for 2018. *Am. Health Drug Benefits* **2018**, *11*, 48–54.

3. Tison, G.H.; Sanchez, J.M.; Ballinger, B.; Singh, A.; Olgin, J.E.; Pletcher, M.J.; Vittinghoff, E.; Lee, E.S.; Fan, S.M.; Gladstone, R.A.; et al. Passive detection of atrial fibrillation using a commercially available smartwatch. 2018. *JAMA Cardiol.* **2018**, *3*, 409–416. [[CrossRef](#)]
4. Pramanik, P.K.D.; Upadhyaya, B.K.; Pal, S.; Pal, T. Internet of things, smart sensors, and pervasive systems: Enabling connected and pervasive healthcare. In *Healthcare Data Analytics and Management*; Dey, N., Ashour, A.S., Bhatt, C., James Fong, S., Eds.; Academic Press: Cambridge, MA, USA, 2019; pp. 1–58. [[CrossRef](#)]
5. Watson, K.; Wells, J.; Sharma, M.; Robertson, S.; Dascanio, J.; Johnson, J.W.; Davis, R.E.; Nahar, V.K. A survey of knowledge and use of telehealth among veterinarians. *BMC Vet. Res.* **2019**, *15*, 474. [[CrossRef](#)]
6. Pacis, D.M.M.; Subido, E.D.C.; Bugtai, N.T. Trends in telemedicine utilizing artificial intelligence. *AIP Conf. Proc.* **2018**, *1933*, 040009. [[CrossRef](#)]
7. Kour, H.; Patison, K.P.; Corbet, N.J.; Swain, D.L. Validation of accelerometer use to measure suckling behaviour in Northern Australian beef calves. *Appl. Anim. Behav. Sci.* **2018**, *202*, 1–6. [[CrossRef](#)]
8. den Uijl, I.; Gómez Álvarez, C.B.; Bartram, D.; Dror, Y.; Holland, R.; Cook, A. External validation of a collar-mounted triaxial accelerometer for second-by-second monitoring of eight behavioural states in dogs. *PLoS ONE* **2017**, *12*, e0188481. [[CrossRef](#)] [[PubMed](#)]
9. Belda, B.; Enomoto, M.; Case, B.C.; Lascelles, B.D.X. Initial evaluation of PetPace activity monitor. *Vet. J.* **2018**, *237*, 63–68. [[CrossRef](#)] [[PubMed](#)]
10. Weiss, G.M.; Nathan, A.; Kropp, J.B.; Lockhart, J.W. WagTag: A dog collar accessory for monitoring canine activity levels. In Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication, UbiComp '13 Adjunct, Zurich, Switzerland, 8–12 September 2013; ACM Press: New York, NY, USA, 2013; pp. 405–414. [[CrossRef](#)]
11. Mejia, S.; Duerr, F.M.; Salman, M. Comparison of activity levels derived from two accelerometers in dogs with osteoarthritis: Implications for clinical trials. *Vet. J.* **2019**, *252*, 105355. [[CrossRef](#)] [[PubMed](#)]
12. Westgarth, C.; Ladha, C. Evaluation of an open source method for calculating physical activity in dogs from harness and collar based sensors. *BMC Vet. Res.* **2017**, *13*, 322. [[CrossRef](#)] [[PubMed](#)]
13. Hansen, B.D.; Lascelles, B.D.X.; Keene, B.W.; Adams, A.K.; Thomson, A.E. Evaluation of an accelerometer for at-home monitoring of spontaneous activity in dogs. *Am. J. Vet. Res.* **2007**, *68*, 468–475. [[CrossRef](#)] [[PubMed](#)]
14. Hoffman, C.L.; Ladha, C.; Wilcox, S. An actigraphy-based comparison of shelter dog and owned dog activity patterns. *J. Vet. Behav.* **2019**, *34*, 30–36. [[CrossRef](#)]
15. Kumpulainen, P.; Valldeoriola, A.; Somppi, S.; Törnqvist, H.; Väättäjä, H.; Majaranta, P.; Surakka, V.; Vainio, O.; Kujala, M.V.; Gizatdinova, Y.; et al. Dog activity classification with movement sensor placed on the collar. In Proceedings of the Fifth International Conference on Animal-Computer Interaction, Atlanta, GA, USA, 4–6 December 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–6. [[CrossRef](#)]
16. Brugarolas, R.; Loftin, R.T.; Yang, P.; Roberts, D.L.; Sherman, B.; Bozkurt, A. Behavior recognition based on machine learning algorithms for a wireless canine machine interface. In Proceedings of the 2013 IEEE International Conference on Body Sensor Networks, Cambridge, MA, USA, 6–9 May 2013; pp. 1–5. [[CrossRef](#)]
17. Petrus, S.; Roux, L. Real-Time Behaviour Classification Techniques in Low-Power Animal Borne Sensor Applications. Ph.D. Thesis, Stellenbosch University, Stellenbosch, South Africa, 2019. Available online: <https://scholar.sun.ac.za:443/handle/10019.1/105744> (accessed on 10 January 2020).
18. den Uijl, I.; Gomez-Alvarez, C.; Dror, Y.; Manning, N.; Bartram, D.; Cook, A. *Validation of a Collar-Mounted Accelerometer That Identifies Eight Canine Behavioural States, including Those with Dermatologic Significance*; British Veterinary Dermatology Study Group: Weybridge, UK, 2016; pp. 81–84.
19. Griffies, J.D.; Zutty, J.; Sarzen, M.; Soorholtz, S. Wearable sensor shown to specifically quantify pruritic behaviors in dogs. *BMC Vet. Res.* **2018**, *14*, 124. [[CrossRef](#)] [[PubMed](#)]
20. Ladha, C.; Hammerla, N.; Hughes, E.; Olivier, P.; Ploetz, T. Dog's life. In Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland, 8–12 September 2013. Available online: <https://dl.acm.org/doi/abs/10.1145/2493432.2493519> (accessed on 10 January 2020).
21. Kiyohara, T.; Orihara, R.; Sei, Y.; Tahara, Y.; Ohsuga, A. Activity recognition for dogs based on time-series data analysis. In *Agents and Artificial Intelligence*; Duval, B., van den Herik, J., Loiseau, S., Filipe, J., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 163–184. [[CrossRef](#)]
22. Nuttall, T.; McEwan, N. Objective measurement of pruritus in dogs: A preliminary study using activity monitors. *Vet. Derm.* **2006**, *17*, 348–351. [[CrossRef](#)] [[PubMed](#)]
23. Plant, J.D. Correlation of observed nocturnal pruritus and actigraphy in dogs. *Vet. Rec.* **2008**, *162*, 624–625. [[CrossRef](#)]
24. Morrison, R.; Penpraze, V.; Beber, A.; Reilly, J.J.; Yam, P.S. Associations between obesity and physical activity in dogs: A preliminary investigation. *J. Small Anim. Pract.* **2013**, *54*, 570–574. [[CrossRef](#)]
25. Helm, J.; McBrearty, A.; Fontaine, S.; Morrison, R.; Yam, P. Use of accelerometry to investigate physical activity in dogs receiving chemotherapy. *J. Small Anim. Pract.* **2016**, *57*, 600–609. [[CrossRef](#)] [[PubMed](#)]
26. Twomey, N.; Diethe, T.; Fafoutis, X.; Elsts, A.; McConville, R.; Flach, P.; Craddock, I. A comprehensive study of activity recognition using accelerometers. *Informatics* **2018**, *5*, 27. [[CrossRef](#)]

27. Foerster, F.; Smeja, M.; Fahrenberg, J. Detection of posture and motion by accelerometry: A validation study in ambulatory monitoring. *Comput. Hum. Behav.* **1999**, *15*, 571–583. [[CrossRef](#)]
28. Olsen, A.M.; Evans, R.B.; Duerr, F.M. Evaluation of accelerometer inter-device variability and collar placement in dogs. *Vet. Evid.* **2016**, *1*, 2–9. [[CrossRef](#)]
29. Martin, K.W.; Olsen, A.M.; Duncan, C.G.; Duerr, F.M. The method of attachment influences accelerometer-based activity data in dogs. *BMC Vet. Res.* **2017**, *13*, 48. [[CrossRef](#)]
30. Aich, S.; Chakrabort, S.; Sim, J.-S.; Jang, D.-J.; Kim, H.-C. The design of an automated system for the analysis of the activity and emotional patterns of dogs with wearable sensors using machine learning. *Appl. Sci.* **2019**, *9*, 4938. [[CrossRef](#)]
31. Pet Insight Project. Available online: <https://www.petinsight.co> (accessed on 16 December 2019).
32. Chambers, R.D.; Yoder, N.C. FilterNet: A many-to-many deep learning architecture for time series classification. *Sensors* **2020**, *20*, 2498. [[CrossRef](#)] [[PubMed](#)]
33. Friard, O.; Gamba, M. BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. *Methods Ecol. Evol.* **2016**, *7*, 1325–1330. [[CrossRef](#)]
34. Hammerla, N.Y.; Plötz, T. Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan, 7–11 September 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1041–1051. [[CrossRef](#)]
35. Gerencsér, L.; Vásárhelyi, G.; Nagy, M.; Vicsek, T.; Miklósi, A. Identification of behaviour in freely moving dogs (*Canis familiaris*) using inertial sensors. *PLoS ONE* **2013**, *8*, e77814. [[CrossRef](#)] [[PubMed](#)]
36. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc: New York, NY, USA, 2019; pp. 8024–8035. Available online: <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf> (accessed on 6 December 2019).
37. Amazon EC2-P2 Instances. Amazon Web Services, Inc. Available online: <https://aws.amazon.com/ec2/instance-types/p2/> (accessed on 6 December 2019).
38. Haghghi, S.; Jasemi, M.; Hessabi, S.; Zolanvari, A. PyCM: Multiclass confusion matrix library in Python. *JOSS* **2018**, *3*, 729. [[CrossRef](#)]
39. Steyerberg, E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*; Springer: New York, NY, USA, 2009. [[CrossRef](#)]