



Supervised machine learning aided behavior classification in pigeons

Neslihan Wittek¹ · Kevin Wittek² · Christopher Keibel³ · Onur Güntürkün¹

Accepted: 16 May 2022 / Published online: 14 June 2022
© The Author(s) 2022

Abstract

Manual behavioral observations have been applied in both environment and laboratory experiments in order to analyze and quantify animal movement and behavior. Although these observations contributed tremendously to ecological and neuroscientific disciplines, there have been challenges and disadvantages following in their footsteps. They are not only time-consuming, labor-intensive, and error-prone but they can also be subjective, which induces further difficulties in reproducing the results. Therefore, there is an ongoing endeavor towards automated behavioral analysis, which has also paved the way for open-source software approaches. Even though these approaches theoretically can be applied to different animal groups, the current applications are mostly focused on mammals, especially rodents. However, extending those applications to other vertebrates, such as birds, is advisable not only for extending species-specific knowledge but also for contributing to the larger evolutionary picture and the role of behavior within. Here we present an open-source software package as a possible initiation of bird behavior classification. It can analyze pose-estimation data generated by established deep-learning-based pose-estimation tools such as DeepLabCut for building supervised machine learning predictive classifiers for pigeon behaviors, which can be broadened to support other bird species as well. We show that by training different machine learning and deep learning architectures using multivariate time series data as input, an F1 score of 0.874 can be achieved for a set of seven distinct behaviors. In addition, an algorithm for further tuning the bias of the predictions towards either precision or recall is introduced, which allows tailoring the classifier to specific needs.

Keywords Actions · Sequence analysis · Ethogram · DeepLabCut · Birds · Deep learning

Introduction

Neuroscience has taken a breathtaking ascent within just a few decades (Cobb, 2020). Despite countless success stories at the molecular, cellular, and clinical levels, the explanation of behavior by reverse engineering of neural components or by other bottom-up means has fallen short (Peebles & Cooper, 2015; Jonas & Kording, 2017). Instead, behavior

itself has also to be analyzed with the same painstaking accuracy as done in other neuroscientific fields (Tinbergen, 1963; Krakauer et al., 2017). As concisely phrased by Jerry Hirsh, “Nothing in neurobiology makes sense, except in the light of behavior” (Hirsh, 1986).

The detailed analysis of animal behavior also paved the way to modern experimental psychology and still greatly contributes to various psychological insights (Thorndike, 1898; Pavlov, 1927; Skinner, 1938; Kilian et al., 2003; Vallortigara et al., 2005; Zentall et al., 2013; Du et al., 2016; Anselme, 2021). Ecology-driven research fields are particularly interested in the evolutionary roots of animal behavior, which can be affected by external factors such as limitations of nutrients, territories, or mates (Brown, 1969; Baker, 1972; Gill & Wolf, 1975; Aragón et al., 2003; Arak, 1983; Bailey, 2003; Brown et al., 2006; Bentsen et al., 2006; Anselme & Güntürkün, 2019), whereas experimental and molecular biology combine their methods with behavioral observations to investigate medical conditions such as Parkinson’s disease and early life stress (Kravitz et al., 2010;

Neslihan Wittek and Kevin Wittek contributed equally to this work.

✉ Neslihan Wittek
neslihan.wittek@rub.de

¹ Faculty of Psychology, Department of Biopsychology, Institute of Cognitive Neuroscience, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany

² Faculty of Mathematics, Computer Science and Natural Sciences, Department of Computer Science, RWTH Aachen University, Aachen, Germany

³ Institute for Internet Security, Westphalian University of Applied Sciences, Gelsenkirchen, Germany

Mundorf et al., 2020). In addition, neuroscientists co-analyze behavioral paradigms in their experimental designs to identify the functional relevance of their neurobiological findings (Miri et al., 2017; Caggiano et al., 2018; Branco & Redgrave, 2020; Packheiser et al., 2021).

In order to quantify animal movement and behavior, both natural-habitat and laboratory experiments have continuously benefitted from on-site manual behavioral observations (von Frisch, 1967; Lindburg, 1969; Gallup, 1970; Calhoun, 1970; Altmann, 1974; Anschel & Talmage-Riggs, 1977; Pepperberg et al., 1995; Pollok et al., 2000; Reiss & Marino, 2001; Dally et al., 2006; Prior et al., 2008). Despite all those pioneering contributions, certain challenges and disadvantages follow in the footsteps of manual behavioral observations: They are not only time-consuming and labor-intensive but also have a grain of subjectivity, which might lead to difficulties in reproducing the experiments (Dell et al., 2014). The issues resulting from subjectivity may be mitigated by using camera video recording systems. Unlike a direct observation, a video recording ensures the capture of complete (within the captured dimensions) and detailed behavioral patterns during the observation period (Tosi et al., 2006). However, analyzing video recordings using a traditional approach involving pencil, paper, and stopwatch is time consuming as well. Furthermore, missed detections are still possible due to fluctuating attention of the observer (Anderson & Perona, 2014; Gomez-Marin et al., 2014; Arac et al., 2019).

Besides all these challenges in analyzing behavior, it should not be forgotten that behavior in itself is a complex, dynamic, and multi-dimensional domain (Gomez-Marin et al., 2014), which makes exploring innovative approaches a sensible strategy. At this stage, the recent technological developments in the field of computer vision in conjunction with a newfound interest in artificial intelligence applications have been supporting researchers: Less time and effort is needed to produce precise datasets of animal movement and behavior and the tracking of the animals can be done automatically, which minimizes the amount of human labor and the potential for missed detections (Dell et al., 2014; Bello-Arroyo et al., 2018). Consequently, researchers have been working with different commercial-proprietary and open-source software for the automated analysis of animal behavior with a focus on a particular animal model or different sets of animal groups (commercial-proprietary-EthoVision: Noldus et al., 2001, VideoTrack: ViewPoint Behavior Technology, ANY-maze: Stoelting, Wood Dale, IL, USA; open-source- SwisTrack: Lochmatter et al., 2008; Ethowatcher: Crispim Junior et al., 2012; JAABA: Kabra et al., 2013; idTracker: Pérez-Escudero et al., 2014; DeepLabCut: Mathis et al., 2018; MouBeAT: Bello-Arroyo et al., 2018; UMATracker: Yamanaka & Takeuchi, 2018; Tracktor: Sridhar et al., 2019; TRex: Walter & Cousin,

2020). These software products have not only provided the foundation for quantitative and precise results like velocity, body-orientation, trajectory, and time spent in a particular area (Evans et al., 2015; Singh et al., 2016; Dankert et al., 2009; Luyten et al., 2014, Wittek et al., 2021), but have also established the groundwork for automated analysis of measuring complex behaviors such as anxiety, stress, aggressiveness, risk assessment, shoaling, and spatial learning in different animal groups (Rodríguez et al., 2004; Choy et al., 2012; Piato et al., 2011; Green et al., 2012; Miller & Gerlai, 2012; Nema et al., 2016; Peng et al., 2016; Mazur-Milecka & Ruminski, 2017; Mundorf et al., 2020). Extending those applications to birds is advisable not only for extending the species-specific knowledge but also for contributing to the bigger picture of the evolutionary process. Most importantly, studies on pigeons have a long tradition in experimental psychology and importantly have contributed to insights about learning and memory (Vaughan & Greene, 1984; Troje et al., 1999; Fagot & Cook, 2006; Pearce et al., 2008; Wilzeck et al., 2010; Rose et al., 2009; Scarf et al., 2016; Güntürkün et al., 2018; Packheiser et al., 2019). However, so far, the application of these automated analyses on birds is still limited. But there is a different species, which lends itself to a closer investigation with regards to automated behavior classification: *Homo sapiens*. Existing applications from industry and academia in the domain of human–computer interaction and computer vision have spun up a vast array of literature, mathematical models, and software approaches for human activity recognition, which should be further investigated in order to establish a baseline.

The increasingly large amount of data acquired by different technical devices and sensors, some of them ubiquitous to today's human life (e.g., "smart devices" such as phones and watches), resulted in an explorative renaissance of machine learning methods by leveraging image- as well as sensor-data (raw or pre-processed) for human activity recognition. Various classification algorithms such as support-vector machine, hidden Markov model, decision tree, random forest, k-nearest neighbors, logistic regression, and stochastic gradient descent have been used to successfully analyze and classify human physical activity (Mannini & Sabatini, 2010; Anguita et al., 2012; Paul & George, 2015; Kolekar & Dash, 2016; Xu et al., 2017; Nematallah et al., 2019; Baldominos et al., 2019). In addition to these traditional machine learning methods, the emergence and widespread availability of new hardware allowing the use of deep learning architectures has motivated a tendency towards using deep learning approaches for human activity recognition as well. These include recurrent neural networks (RNNs) (Murakami & Taguchi, 1991; Murad & Pyun, 2017; Carfi et al., 2018; Koch et al., 2019), long short-term memory (LSTM) (Chen

et al., 2016; Singh et al., 2017; Pienaar & Malekian, 2019) and convolutional neural networks (CNN) (Wang et al., 2019; Lee et al., 2017; Gholamrezaei & Taghi Almodarresi, 2019; Naqvi et al., 2020; Cruciani et al., 2020; Mehmood et al., 2021; Mekruksavanich & Jitpattanakul, 2021).

In light of this, the current study aims to compound the technical knowledge acquired in both human and non-human domains in order to establish automated bird behavior classification techniques. We used DeepLabCut (DLC: Mathis et al., 2018) as a markerless pose estimation tool to procure multivariate time series data. As a further step, we developed a module named Winkie (a name which was inspired by a Dickin Medal owning pigeon of the same name that had assisted in the rescue of an aircrew during the Second World War). This module consists of submodules for pre-processing and normalizing of the DLC data. Afterward, we applied and compared different machine learning and deep learning architectures to classify pigeon behaviors like eating, standing, walking, head shaking, tail shaking, preening, and fluffing. As a machine learning architecture, random forest gave a high weighted F1 score (0.81) over all behaviors and showed good performance for behaviors that were stable along spatial and temporal dimensions (such as eating,

fluffing, preening, standing). The deep learning architecture InceptionTime, as a one-dimensional convolutional neural network (CNN), also demonstrated high overall performance (0.87). However, the particular performance for the highly dynamic behaviors such as head shake and tail shake were increased substantially in comparison to random forest.

Method

Using the Winkie module as part of a research workflow consists of a sequential multi-step process as shown in Fig. 1A. The same process is used for evaluating its performance itself and each step will be discussed in the further sections.

Data acquisition and manual observation

Eight naïve adult homing pigeons (*Columba livia*) from local breeders were maintained at 85–90% of their free-feeding body weight throughout the experiment, while water was accessible ad libitum. The experiment was conducted

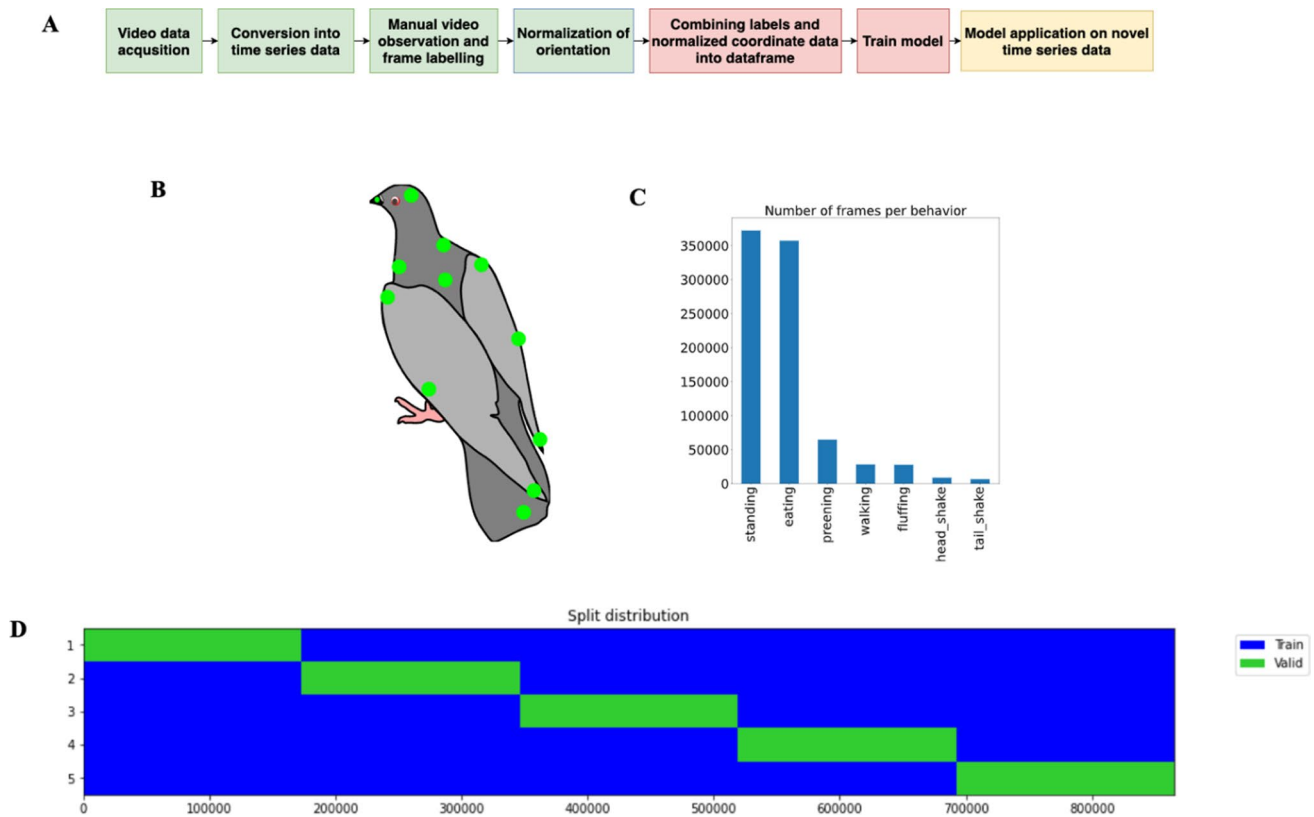


Fig. 1 Data preparation. **A** The Winkie module consists of a sequential multi-step process for pigeons. **B** The tracked body points (head, beak, left-right neck, body, left up-middle-down wing, right up-

middle-down wing, tail). **C** The number of frames per behaviors extracted after applying LabelAssistant. **D** The split distribution of fivefold non-shuffle cross-validation

in a wooden box with a feeder located in the middle part. All procedures followed the German guidelines for the care and use of animals in science and were in accordance with the European Communities Council Directive 86/609/EEC concerning the care and use of animals for experimentation. They were also approved by Ruhr University Bochum, Germany.

Depending on their activity level in the experiment, pigeons received between 10 and 20 (10 min each) sessions in which they could freely move and consume grains. For instance, while highly active individuals were trained for 20 sessions to increase the possibility of dynamic behavior occurrence, the stable individuals received ten sessions. Each session was recorded with a GoPro HERO7 at 119.88 frames per second with resolution of 1280 x 960 pixels. Initially, the videos were manually checked in detail to detect any occurrence of individual behaviors as described below and shown in [Supplementary Video1](#):

Eating: Consuming the food in the feeder or food dropped on the experiment platform.

Standing: Remaining at the same location for an indeterminate period.

Walking: Changing the location.

Head Shaking: Moving the head along a curve in fluent and repeated motion.

Tail Shaking: Moving the tail along a curve in fluent and repeated motion.

Preening: Maintenance behavior that involves the use of the beak to reposition feathers on different parts of the body. A preening event started each time when the beak touched the body and finished once the beak lost contact with the body.

Fluffing: Partial or total extension of one or both wings and ruffling feathers. Additional flapping of wings might occasionally occur.

The observer noted the time slice (starting and ending timecode) in which the behaviors occurred.

Pre-processing video data

Markerless pose estimation and manual observation verification

Video-tracking was performed using the machine-learning-based tracking software DeepLabCut (DLC: Mathis et al., 2018). Since we were interested in behaviors in which different body parts are actively involved, we tracked different points from the pigeon body as shown in Fig. 1B (head, beak, left-right neck, body, left up-middle-down wing, right up-middle-down wing, tail). The data acquired via DLC

processing consist of multivariate time series data, which is a series of location values per body part over a period of time. For instance, a 10-min video recording results in approximately 71,928 frames (600 s x 119.88 fps). In this sense, our usage of DLC can be understood as a lossfull, but semantically enriched and transformed, data reduction step: For each frame, we go from raw (decompressed) 1280 x 960 x 8 bit = 9.5 MBit to 10-body parts x 2 x 32 bit = 640 bit. This comes down to a reduction factor on the bandwidth of the input data of roughly 15,000. In total, this process generated multivariate time series data for 10,424,241 frames.

In order to ensure the feasibility of detecting behaviors exclusively from the multivariate time series data, it was necessary to check whether a human observer can identify the behaviors mentioned above on the DLC as well. Therefore, we developed a module called *Pigeon Animator* for visualization of time slices with frame precision. In addition to verifying the behaviors and performance of the tracking, we narrowed the time slices by replacing the timecodes with frame numbers (see [Supplementary Video2](#)).

Labeling

After applying the verified labels of the manual observer to the individual frame data, non-labeled frames were removed from the data set, which led to 865,548 remaining labeled frames. In order to make this process less error-prone, we developed a custom module *LabelAssistant* that would ensure the integrity of the labels with regards to the DLC output and safeguard against specific error classes (e.g., ensuring consistency in the label names). As shown in Fig. 1C, we ended up with an imbalanced data set, especially since the natural frequency of occurrence of the behaviors is already unbalanced. The challenge of class imbalance will be discussed in the following sections in more detail.

Transformation and normalization of two-dimensional Euclidean input data

DLC tracks absolute coordinates, while the behavior should be considered by looking at the data from a relative standpoint in order to simplify classification and pattern matching (e.g., for hand gesture analysis: Do et al., 2020). In *DLCImporter*, we developed parameterizable functions for pre-processing the DLC data by normalizing it such that the body is translated into the origin (and the other parts translated accordingly, thereby representing the relative position to the body) while rotating all points such that the vector between body and the middle of the neck (basically the “spine”) becomes parallel to the y-axis. Since the body is in the origin, the spine is thereby implicitly located along the y-axis.

For each frame, the displacement vector s is defined as:

$$s = (-\text{coordinate}(\text{body}, x), -\text{coordinate}(\text{body}, y)) \quad (1)$$

and each body part bp is translated using s as the translation vector in the translation function T_V :

$$T_V(bp) = bp + s \quad (2)$$

In addition, a new body part *middle neck* is added which for our data is defined as:

$$v_{\text{middle neck}} = v_{\text{left neck}} - v_{\text{right neck}} \quad (3)$$

Based on this vector, the necessary rotation rot_{norm} as the angle in degree between the positive x -axis and the vector $v_{\text{middle neck}}$ is calculated:

$$rot_{norm} = \arctan2(v_{\text{middle neck } x}, v_{\text{middle neck } y}) \times \frac{180}{\pi} \quad (4)$$

Using this angle, the rotation matrix R_{norm} was constructed and applied on all body parts:

$$R_{norm} = \begin{bmatrix} \cos rot_{norm} & -\sin rot_{norm} \\ \sin rot_{norm} & \cos rot_{norm} \end{bmatrix} \quad (5)$$

Machine learning and deep learning architectures for behavioral classification

As discussed in the Introduction section, there is a plethora of machine learning and deep learning architectures that can be used to classify human activity and behaviors, each with its own strength and weaknesses for specific permutations of domains and input data. In addition, the field is constantly evolving, with new architectures and improved methods emerging continuously. In order to demonstrate the general feasibility of our approach, we selected *random forests* as a machine learning architecture and *InceptionTime* as a deep learning architecture.

The raw features that were used in both *random forest* and *InceptionTime* were defined according to their tracking likelihood values as given by DLC, which was an indication of the overall stability of the tracked body point. Accordingly, the x and y pixel coordinates of ‘head, left-right neck, body, left up-middle-down wing, right up-middle-down wing, tail’ were used as features, while ‘beak’ was excluded since tracking was unstable due to the frequent occlusion of the beak by the pigeon itself.

Overall generalization performance of fitted models is measured using five-fold non-shuffled cross-validation-score (Fig. 1D) as the arithmetic mean (10) of the weighted F1 score (9), to cater for imbalances of classes in the data set. F1 score (8) is an established scoring mechanism for measuring the accuracy of an information retrieval system and

is defined as the harmonic mean of precision (6) and recall (7) (Rijsbergen, 1979; Chinchor, 1992). Although most of the literature studies have opted to shuffle data as part of the train-test split, due to the time series nature of the data at hand containing implicit dependencies between consecutive data points, we decided against it, since this would lead to unrealistically good test scores (since very similar data can end up in the train- and test-set). This problem is further amplified by the fact that the original video recordings were performed using a high framerate of 119.8 FPS. In addition, the last fold (first 80% as training-set and last 20% as test-set) was used to show the classification performance of individual behaviors.

$$\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (6)$$

$$\text{recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (7)$$

$$F_1 = \frac{2}{\text{precision}^{-1} + \text{recall}^{-1}} \quad (8)$$

$$F1_{\text{weighted}} = \frac{1}{\text{total_targets}} \times \sum_{i=1}^{n_classes} F1_i \times \text{targets}_i \quad (9)$$

$$cv_5 = \frac{1}{5} \left(\sum_{i=1}^5 F1_{\text{weighted}}(i) \right) \quad (10)$$

All evaluations were performed using an AMD Ryzen 9 5950X @ 3.4–4.9 GHz, 32GB RAM, NVIDIA GeForce RTX 2070 Super 8GB RAM, running on Microsoft Windows 10 Pro Build 19043. The Python machine-learning library scikit-learn (Pedregosa et al., 2011) was used for the random forest classifier and overall performance metrics while the deep learning stack of tsai, fast.ai and PyTorch was used for InceptionTime (Paszke et al., 2019; Oguiza, 2020; Howard & Gugger, 2020).

Decision Trees and Ensemble Methods (Random Forest)

A decision tree in the context of machine learning can be understood as a binary tree with each node in the tree splitting the source set based on an inferred criteria of an input feature, leading to leaves containing the resulting class or a specific probability distribution of the classes. Decision trees used as classification trees have been shown to be an intuitive way to classify and label objects and if they are trained on high-quality data, they ensure very accurate predictions (Caruana & Niculescu-Mizil, 2006; Kingsford &

Salzberg, 2008). The intuitive character of decision trees can be demonstrated by giving an example of how a human might intuitively build one: For example, if you want to construct a decision tree to identify the owner of a chirping sound, you can narrow down the possible answers by asking several consecutive and potentially dependent questions for binary splitting: Which birds are abundant during the current season, which ones are songbirds, is it night or day, etc. Each question will narrow the options and you will go on asking these questions until you reach a highly certain answer. Depending on the data used as input (the features that can be extracted from this input) and the possible answers to a sequence of questions (nodes), the resulting leaf node might contain a clear-cut answer (it is a Jay), or a distribution of class predictions (70% Kookaburra, 30% Lyrebird).

To further improve the predictive performance of machine learning algorithms, ensemble learning can be used to combine multiple different models into a single model (Dietterich, 2000; Peterson & Martinez, 2005). Random forests are an ensemble learning method for classification that makes use of a set of different decision trees and is shown to generally outperform decision trees (Ho, 1995; Pirayonesi & El-Diraby, 2020).

For our model, the number of maximum features per split max_f was defined as:

$$max_f = \sqrt[n_{features}]{} \tag{11}$$

A good hyperparameter value $n_{estimators}$ for the number of trees in the forest was determined by calculating the validation curve for the set $s_{estimators}$:

$$s_{estimators} = \{x \in \mathbb{N} \mid 1 \leq x \leq 100\} \tag{12}$$

A reasonable number of trees considering the tradeoff between accuracy and time efficiency was selected as the hyperparameter to detect the performance of individual behaviors. In addition, the learning curves for different training set sizes were evaluated to determine the correlation between training set size, classification performance, and training time.

The model was created by segmenting the time series data into windows of different sizes using a sliding window approach with a step size of 1. The effect of different sizes on the performance was evaluated using a validation curve with the number of consecutive frames included in the input vector as a hyperparameter:

$$dim_{v_{input}} = (n_{features} \times window_{size}, 1) \tag{13}$$

In order to combat the class imbalance, to which decision trees are sensitive, all models were trained using balanced class weights, with the weight w_c for a class c adjusted to be inversely proportional to class frequencies in the input data (Sun et al., 2009):

$$w_c = \frac{total\ number\ of\ observations}{number\ of\ classes \times number\ of\ observations_c} \tag{14}$$

InceptionTime

Through *AlexNet* winning the visual recognition challenge competition (*ImageNet*) in 2012, deep CNNs have been established as a state-of-the-art technique for domains such as image recognition, object detection or natural language processing, often reaching human levels of performance (Ren et al., 2015, Fawaz et al., 2019). Accordingly, Fawaz et al. (2020) propose *InceptionTime* to be an *AlexNet* equivalent for time series classification, in which an ensemble of deep CNN models (inception modules) is used for classification of multivariate time series data.

The optimal depth of the network depends on the lengths of patterns contained in each time series segment. In order to evaluate the effect of the depth hyperparameter on the model performance, we calculated the validation curve for the parameter range 1–6, with 6 being the default for *InceptionTime*.

Similar to the random forest, the time series was segmented using a sliding window approach with step size of 1. The window size was kept at 16 frames, which seemed suitable to capture not only long patterns, but also sudden and short ones.

The fitting of the models was done using one-cycle super-convergence training for learning rate adaption as dynamic hyperparameter tuning (Smith & Topin, 2018). Mock training with cyclical learning rates was used to determine a good maximum learning rate (Smith, 2017), with the steepest point of the resulting learning rate curve being selected as the maximum learning rate.

According to Smith (2018), although historically small batch sizes have been recommended for regularization effects, when applying a one-cycle learning rate schedule (as we do) a high batch size can be used to minimize computational time, while still achieving high performance. With regards to our available GPU memory, a batch size of 1024 was selected.

Post processing

Applying any of the aforementioned models (on novel or existing data) returns a probability vector x_p with $dim(x_p) = n_{classes}$, where $n_{classes}$ is the total of different classes, for each classified frame, respectively each classified time window. In addition, the sum of all vector elements is always equal to 1. Conservatively, applying:

$$b(f) = \operatorname{argmax}(x_p(f)) \quad (15)$$

will yield the predicted behavior b at frame f .

For binary classification models, traditionally different threshold values for selecting a prediction (compared to 0.5) can be applied to further tune the results

with regards to precision and recall, depending on the needs of the application (Fielding & Bell, 1997). Inspired by these approaches, we propose an algorithm that allows individual thresholds for behavior tuples in a multi-class model:

```

Input
-----

probas <- model output as a list of probability vectors
tuple_thresholds <- list of custom thresholds per class tuple

Return
-----

Decoded prediction classes based on individual thresholds.

Function:
-----

def apply_individual_thresholds(probas, tuple_thresholds):
    tuned_predictions = []
    for p in probas:
        # top 2 probabilities
        top_indices = argpartition(p, -2)[-2:]
        if (top_indices in tuple_thresholds):
            # individual threshold for top tuple
            t = tuple_thresholds[top_indices]
            if p[top_indices[1]] > t:
                # if second highest probability is over
                # threshold, the second element of tuple gets
                # precedence
                tuned_predictions.append(top_indices[1])
            else:
                # not over threshold
                tuned_predictions.append(argmax(p))
        else:
            # no specific threshold for top tuple
            tuned_predictions.append(argmax(p))
    return tuned_predictions

```

Algo 1: Tuned predictions

It is up to the user how those tuples are defined or optimized. However, we will show the effect of some a-posteriori chosen example values in the results section.

Results

Animal tracking

DLC training was performed using 1,030,000 iterations, achieving a root mean square error (RMSE) over all tracked body parts of 2.53 pixels for the train set and 6.41 pixels for the test set. Using a prediction cutoff value of 0.6, the train error remained the same and the test error could be reduced to 6.16 pixels. For our given video resolution of 1280 x 960 pixels, this translates to roughly 5.6 mm in the physical world.

Random forest performance

The validation curve for the $n_{\text{estimators}}$ hyperparameter was calculated and analyzed, revealing a sufficiently good cross-validation score of 0.79 for 20 trees, with the maximum score of 0.81 occurring for 95 trees (Fig. 2A). Based on this

finding, the learning curve for 20 trees was calculated, showing a continuous increase of the cross-validation score as a function of training-set size. However, the learning curve seems to reach a saturation point, for the maximum amount of available training data in our case (Fig. 2B). Further window size evaluation using a validation curve revealed that the overall performance was not strongly affected by the size of the window (Fig. 2C: $F1_{\text{single frame}} = 0.807 \pm 0.054$, $F1_{2 \text{ frames}} = 0.806 \pm 0.042$, $F1_{4 \text{ frames}} = 0.813 \pm 0.038$, $F1_{8 \text{ frames}} = 0.827 \pm 0.038$, $F1_{16 \text{ frames}} = 0.835 \pm 0.031$, $F1_{32 \text{ frames}} = 0.850 \pm 0.031$, $F1_{64 \text{ frames}} = 0.852 \pm 0.043$). Since sudden behaviors occurred in short bursts of roughly 16 frames, we further compared the single frame and 16 frames performance in detail for individual classes as shown in Fig. 3. Both models gave high classification performance for the behaviors that were stable along spatial and temporal dimensions (meaning the behavior can be accurately classified by assessing the posture in a single frame). The individual behaviors' classification performance remained mostly similar, except for preening and walking. While preening detection was slightly increased for 16 frames, walking detection was slightly decreased. Note that our transformation and normalization steps on the input data remove characteristics of the walking movement, since the coordinates are transformed into a more stable position.

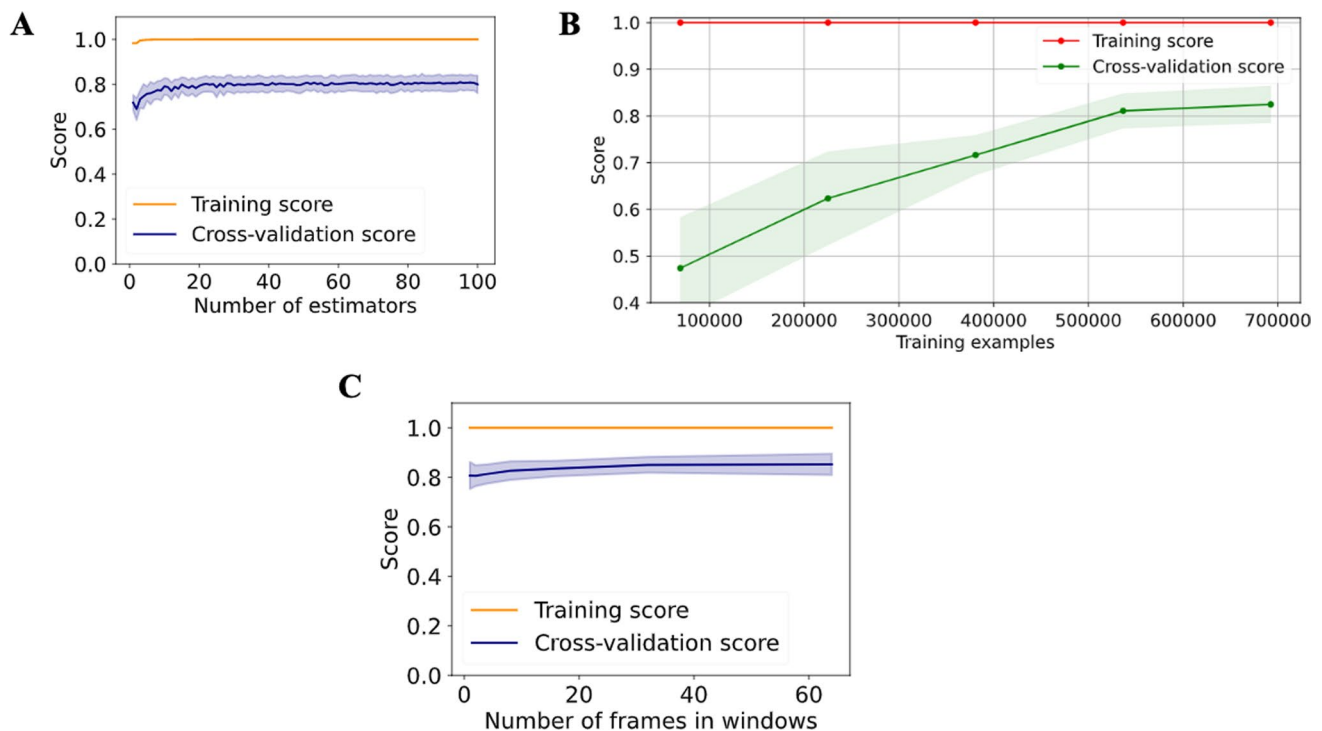


Fig. 2 Random forest evaluation. **A** Validation curve for random forest for number of trees as hyperparameter. No substantial improvement of score for $n > 20$. **B** Learning curve and performance for different amounts of training examples. **C** Validation curve for different window sizes

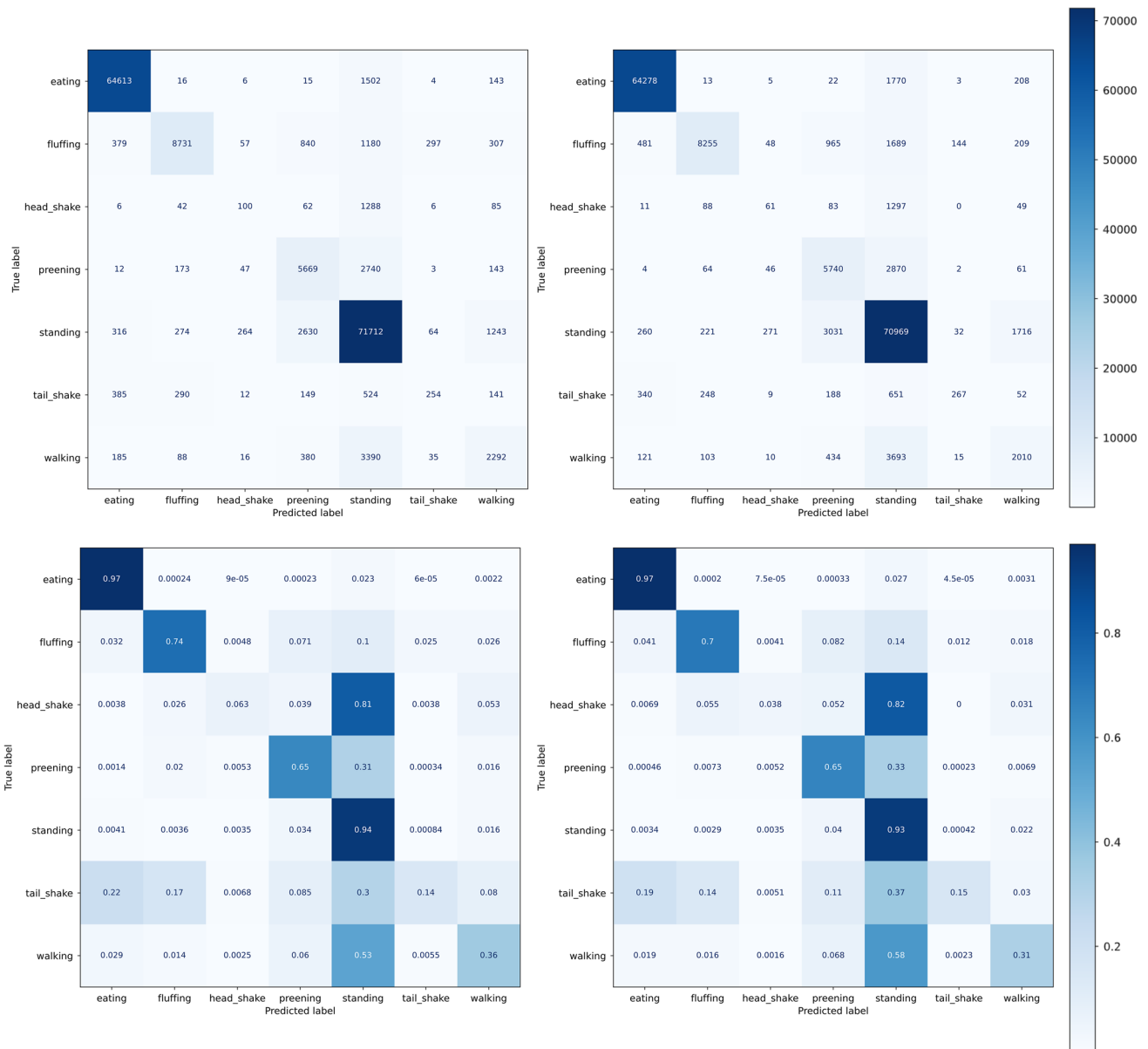


Fig. 3 Confusion matrix for random forest. Confusion matrix for random forest with single frame (left) and sixteen frames (right) windows size (top absolute, bottom relative)

InceptionTime performance

The validation curve for the depth hyperparameter indicated no significant effect of depth on the generalization performance as seen in the cross-validation score (Supplementary Figure 1). In order to reduce the complexity of the model and reduce the potential for overfitting, the smallest depth value of 1 with a F1 cross-validation score of 0.874 ± 0.031 (which was higher than the best scores achieved using *random forest*) was selected for the further evaluation. By calculating the confusion matrix on the last fold, similarly to *random forest* a good performance

was acquired for behaviors that were stable along spatial and temporal dimensions. In addition, an increase in performance, compared to *random forest*, was also achieved on highly dynamic behaviors such as head shake and tail shake (Fig. 4A and Fig. 4B) ($\text{recall}_{\text{head shake}} 0.064$ vs. 0.36 and $\text{recall}_{\text{tail shake}} 0.16$ vs. 0.54).

Post processing and model application

When analyzing the *InceptionTime* confusion matrix, we observed a prevalent confusion between ‘head shake - standing’ and ‘preening - standing’. Based on this observation, we

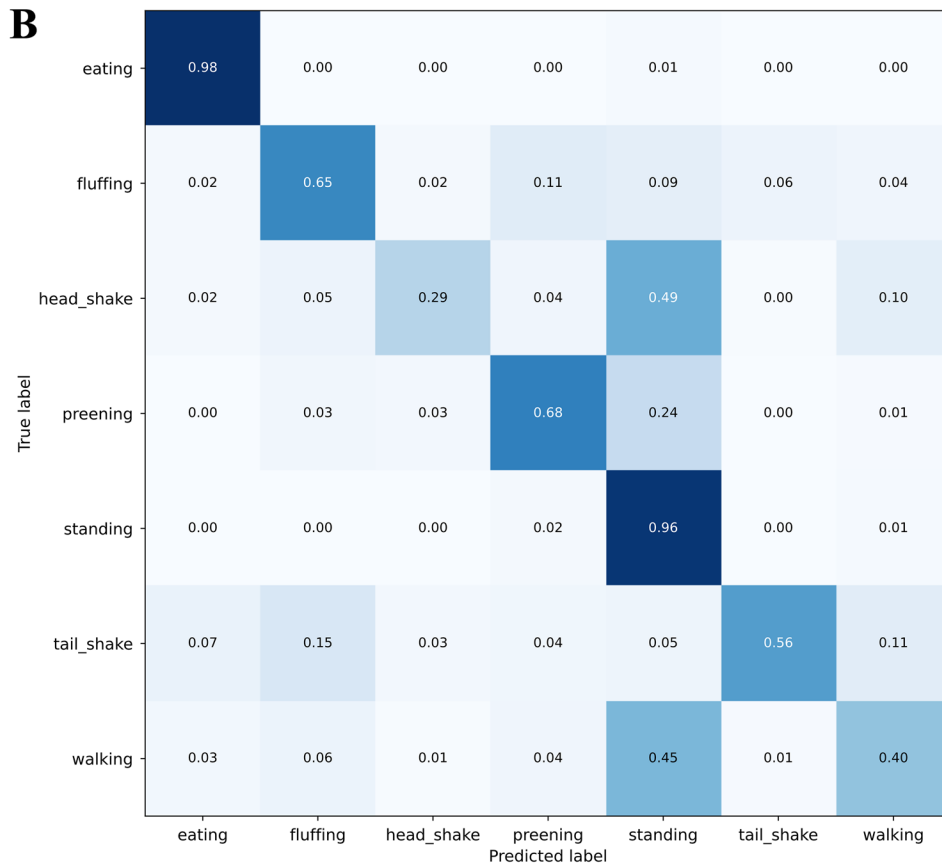
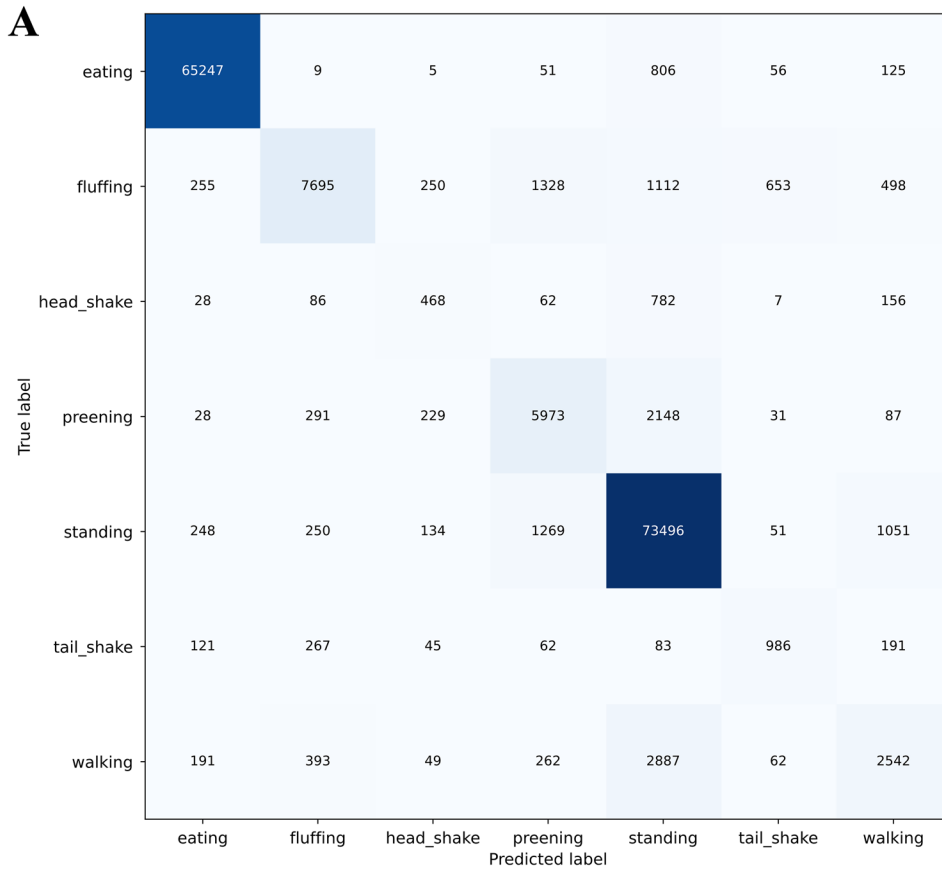


Fig. 4 Confusion matrix for InceptionTime. **A** Confusion matrix for InceptionTime with absolute values. **B** Confusion matrix for InceptionTime with relative values

defined the dynamic tuple thresholds as follows: [(standing, head shake): 0.2, (standing, preening): 0.1]

This changed the predictions in favor of precision (head shake increased to 0.63 and preening increased to 0.78) but led to worse recall (head_shake dropped to 0.08 and preening dropped to 0.57). Therefore, the tuple threshold needs to be adjusted to personal needs, e.g., is it more important to not miss any potential behaviors, or to reduce the number of false positives?

There are different possibilities to evaluate the output of the model for new data. While it is possible to directly work with the model output in a quantitative way, it seems desirable to also acquire forms of visualization that lend itself better to some form of human “quality control”. It is therefore possible to render the original videos with applied predictions (Fig. 5A, Supplementary Video3) or visualize the predicted behaviors over time in the form of an ethogram (Fig. 5B). Both techniques can also be effectively used in conjunction. By assessing the ethogram, a user is able to gather a general overview of the occurring behaviors at a specific point in time at a quick glance. Interesting (or suspicious) looking predictions can be counter-checked using the rendered videos containing the predictions as a text overlay. Especially in combination with tuned tuple thresholds, this can lead to a process that, while not fully automated, significantly augments the previous manual and laborious process.

Discussion

We have demonstrated the feasibility of adopting existing machine learning classification approaches for pigeon behavior by using a simple single camera setup without further tracking equipment. To further improve the usability of our approach, we have developed and released the open-source software library Winkie to act as a starting point for future improvements and developments. Winkie is usable with commercial off-the-shelf computer hardware. While it might be possible to perform the classification directly on the video streams (Bohnslav et al., 2020), our software uses multivariate time series data as created by DLC to reduce the size and complexity of the video input data. Therefore, our software positions itself inside an ecosystem of emerging de facto industry standards of the scientific open-source community. Furthermore, users can configure the software depending on their needs, to change its bias between precision and recall for specific pairs of behaviors.

Although there is a movement among passionate psychologists and neuroscientists to augment their experimental paradigms with automated behavioral tracking (Lochmatter

et al., 2008; Crispim Junior et al., 2012; Kabra et al., 2013; Pérez-Escudero et al., 2014; Mathis et al., 2018; Bello-Arroyo et al., 2018; Yamanaka & Takeuchi, 2018; Sridhar et al., 2019), existing automation tools generally lack support for analyzing bird-specific behaviors. Besides leading to significant amounts of time savings, our software aimed to extend the species-specific knowledge as a catalyst towards the initiation of the classification of bird behavior and to inspire new approaches (Miller, 1988). Considering the challenges to combine behavioral and neurophysiological measurements, we focused on developing a software that can be used with raw behavioral data being captured from simple hardware setups, such as single camera video recordings.

The relevance of our approach becomes visible when concentrating for example on behaviors like preening, scratching, or head shaking, which are typical avian maintenance and comfort actions (Cotgreave & Clayton, 1994). However, encountering a stressful situation induced by a competitor or predator can elevate the occurrence of these activity patterns (Delius, 1967, 1988; Fernández-Juricic et al., 2004; Wittek et al., 2021). Similarly, preening rates also increase after injections of dopamine or adrenocorticotrophic hormone, with the latter also showing increased head shaking (Delius et al., 1976; Delius, 1988; Acerbo, 2001; Kralj-Fiser et al., 2010). Thus, these actions can serve as a behavioral readout of social conflicts and/or neural processes. But although a vast variety of research has reported these behaviors (Miller, 1988; Moyer et al., 2003; Prior et al., 2008; Clary & Kelly, 2016; Kraft et al., 2017; Wittek et al., 2021), there has been no exact classification and automated analysis of them so far. By using our approach, it is easily possible to disambiguate and quantify different kinds of reactions of the animal along the time frame in stressful contexts and/or when injected with various drugs. Thus, we anticipate that this open-source library, and other developments inspired by it, will pave the way for a more quantitative behavioral analysis of different bird species and beyond.

Future directions and challenges

In this manuscript, we demonstrated how machine learning systems can support classical experimental-psychological and ethological approaches by detecting and quantifying avian behavior. It is important to note that future studies should bear in mind that not only the amount but also the sequence of behavior contains highly relevant insights. Besides stereotypical behavioral patterns recorded by classical ethological approach, there are also subpatterns which might be of interest. This can be described with an ontology in which patterns are an aggregate of subpatterns. For example, head shake can be defined as a continuous and alternating sequence of the head-move-left and head-move-right subpatterns that can remain undetected by manual

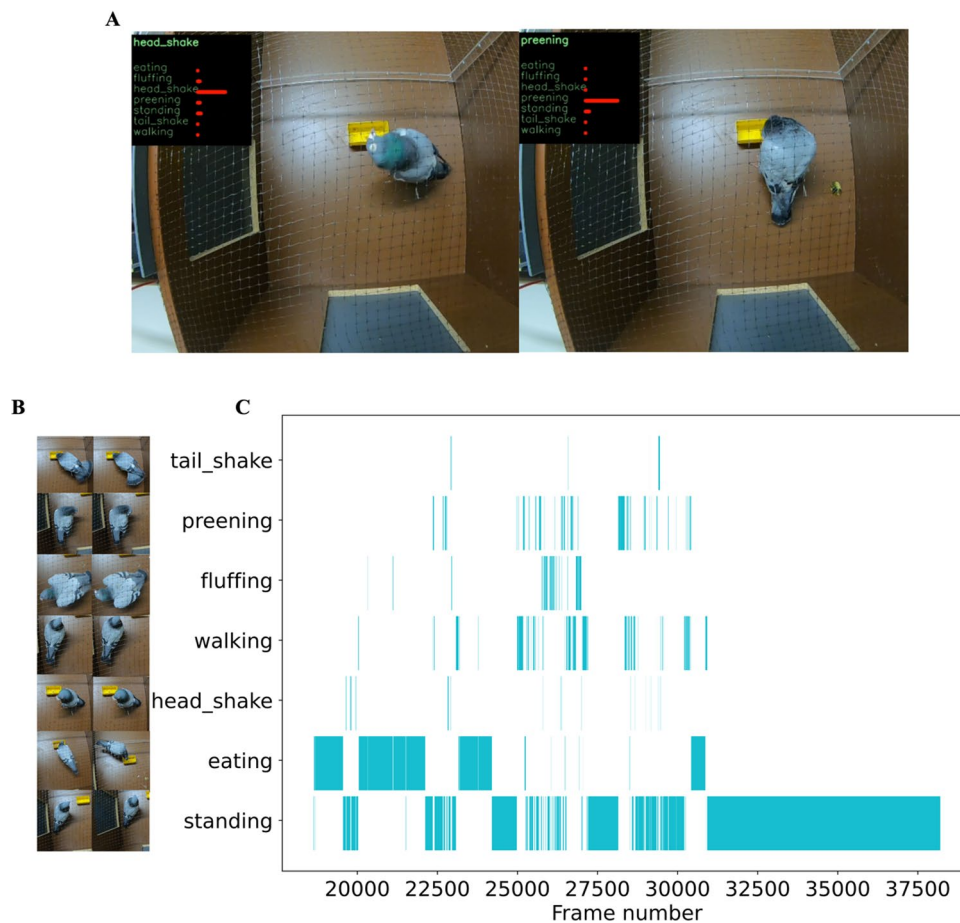


Fig. 5 Classification results on novel video. **A** One possibility to evaluate the model performance is applying the model to the new data to get predictions and render the original videos with the predictions as overlay. **B** Behaviors that were checked in the videos. **C** Ethogram of

predictions for novel data. Observer can go to the related frame number in the original or rendered video to double check the occurrence of the behavior

observation or supervised learning, simply because they are ignored or unknown (Luxem et al., 2020). One promising solution to detect such subpatterns and possible behavioral sequences is using unsupervised machine learning techniques. Besides, in order to understand the functional framework of these behavioral sequences, their correlation with neural activity patterns is easy to implement. In addition, for the behaviors that are hard to capture in two dimensions due to occlusions resulting from the camera perspective, three-dimensional tracking should be considered (Nath et al., 2019).

We have used the DLC output as-is, without applying the filters available in DLC or implementing our own. The lack of filtering might lead to glitches in tracking and anatomically impossible movements. Besides using the available DLC filters or other generic approaches for smoothing such as Kalman filtering (Kalman, 1960) or applying the Ramer–Douglas–Peucker algorithm (Wu & Marquez, 2003), tracking can be also smoothed and aliased by formulating

anatomical constraints for the tracked skeleton through inverse and forward kinematics (Halvorsen et al., 2008; Nilsson et al., 2020).

Besides our multiclass classification approach, applications from the human domain have shown promising results when using fewer classes or ensemble classifiers with multiple binary classification models (Jethanandani et al., 2019), which possibly induces better performance. As we explained in the Method section, we ended up with an imbalanced data set in which rare behaviors like head shake and tail shake were not equally present. Although this fits the natural occurrence frequency of these behaviors, generating a more balanced dataset, by including more data of minority classes, undersampling of majority classes or by using synthetic oversampling techniques on the minority classes (such as SMOTE: Chawla et al., 2002), could lead to better performance for all variants of classification.

Overall, we demonstrated that existing machine learning approaches can be used in conjunction with markerless

pose-estimation tracking data in pigeons – a classic laboratory animal in psychological research on learning, memory, and cognition. The trained model showed high performance on the validation data that was never seen by the model before. In addition, we developed an open-source library as a starting point for further automated classification of bird behaviors. Our system is interface compatible with other machine learning architectures from scikit-learn and PyTorch and is thereby naturally extensible. We are hopeful that our system will help other scientists to extract detailed behavioral data under all kinds of different experimental conditions.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13428-022-01881-w>.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the German Research Council (DFG) through SFB1372 (project number 395940726), project Neu06 to OG.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Acerbo, M. (2001). *The role of dopamine and glutamate in associative learning by the pigeon (Columba livia)* (dissertation). Mathematics and Natural Sciences Faculty of University of Konstanz, Konstanz, Germany.
- Altmann, J. (1974). Observational Study of Behavior: Sampling Methods. *Behavior*, *49*(3/4), 227–267.
- Anderson, D. J., & Perona, P. (2014). Toward a Science of Computational Ethology. *Neuron*, *84*(1), 18–31. <https://doi.org/10.1016/j.neuron.2014.09.005>
- Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2012). Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine. In J. Bravo, R. Hervás, & M. Rodríguez (Eds.), *Ambient Assisted Living and Home Care* (pp. 216–223). Springer. https://doi.org/10.1007/978-3-642-35395-6_30
- Anschel, S., & Talmage-Riggs, G. (1977). Social organization of captive monandrous squirrel monkey groups (*Saimiri sciureus*). *Folia Primatologica. International Journal of Primatology*, *28*(3), 203–215. <https://doi.org/10.1159/000155810>
- Anselme, P. (2021). Effort-motivated behavior resolves paradoxes in appetitive conditioning. *Behavioural Processes*, *193*, 104525. <https://doi.org/10.1016/j.beproc.2021.104525>
- Anselme, P., & Güntürkün, O. (2019). Incentive hope: A default psychological response to multiple forms of uncertainty. *The Behavioral and Brain Sciences*, *42*, e58. <https://doi.org/10.1017/S0140525X18002194>
- Arac, A., Zhao, P., Dobkin, B. H., Carmichael, S. T., & Golshani, P. (2019). DeepBehavior: A Deep Learning Toolbox for Automated Analysis of Animal and Human Behavior Imaging Data. *Frontiers in Systems Neuroscience*, *13*. <https://doi.org/10.3389/fnsys.2019.00020>
- Aragón, P., López, P., & Martín, J. (2003). Differential Avoidance Responses to Chemical Cues from Familiar and Unfamiliar Conspecifics by Male Iberian Rock Lizards (*Lacerta monticola*). *Journal of Herpetology*, *37*(3), 583–585.
- Arak, A. (1983). Sexual selection by male–male competition in natterjack toad choruses. *Nature*, *306*(5940), 261–262. <https://doi.org/10.1038/306261a0>
- Bailey, W. J. (2003). Insect duets: Underlying mechanisms and their evolution. *Physiological Entomology*, *28*(3), 157–174. <https://doi.org/10.1046/j.1365-3032.2003.00337.x>
- Baker, R. R. (1972). Territorial Behavior of the Nymphalid Butterflies, *Aglais urticae* (L.) and *Inachis io* (L.). *Journal of Animal Ecology*, *41*(2), 453–469. <https://doi.org/10.2307/3480>
- Baldominos, A., Cervantes, A., Saez, Y., & Isasi, P. (2019). A Comparison of Machine Learning and Deep Learning Techniques for Activity Recognition using Mobile Devices. *Sensors (Basel, Switzerland)*, *19*(3). <https://doi.org/10.3390/s19030521>
- Bello-Arroyo, E., Roque, H., Marcos, A., Orihuel, J., Higuera-Matas, A., Desco, M., Caiolfa, V. R., Ambrosio, E., Lara-Pezzi, E., & Gómez-Gavero, M. V. (2018). MouBeAT: A New and Open Toolbox for Guided Analysis of Behavioral Tests in Mice. *Frontiers in Behavioral Neuroscience*, *12*. <https://doi.org/10.3389/fnbeh.2018.00201>
- Bentsen, C. L., Hunt, J., Jennions, M. D., & Brooks, R. (2006). Complex multivariate sexual selection on male acoustic signaling in a wild population of *Teleogryllus commodus*. *The American Naturalist*, *167*(4), E102–E116. <https://doi.org/10.1086/501376>
- Bohnslav, J. P., Wimalasena, N. K., Clausing, K. J., Yarmolinsky, D., Cruz, T., Chiappe, E., Orefice, L. L., Woolf, C. J., & Harvey, C. D. (2020). DeepEthogram: A machine learning pipeline for supervised behavior classification from raw pixels. *BioRxiv*. <https://doi.org/10.1101/2020.09.24.312504>
- Branco, T., & Redgrave, P. (2020). The Neural Basis of Escape Behavior in Vertebrates. *Annual Review of Neuroscience*, *43*(1), 417–439. <https://doi.org/10.1146/annurev-neuro-100219-122527>
- Brown, J. L. (1969). Territorial Behavior and Population Regulation in Birds: A Review and Re-Evaluation. *The Wilson Bulletin*, *81*(3), 293–329.
- Brown, W. D., Smith, A. T., Moskalik, B., & Gabriel, J. (2006). Aggressive contests in house crickets: Size, motivation and the information content of aggressive songs. *Animal Behavior*, *72*(1), 225–233. <https://doi.org/10.1016/j.anbehav.2006.01.012>
- Caggiano, V., Leiras, R., Goñi-Erro, H., Masini, D., Bellardita, C., Bouvier, J., Caldeira, V., Fisone, G., & Kiehn, O. (2018). Mid-brain circuits that set locomotor speed and gait selection. *Nature*, *553*(7689), 455–460. <https://doi.org/10.1038/nature25448>
- Calhoun, J. B. (1970). Population density and social pathology. *California Medicine*, *113*(5), 54.
- Carfi, A., Motolese, C., Bruno, B., & Mastrogiovanni, F. (2018). Online Human Gesture Recognition using Recurrent Neural Networks and Wearable Sensors. *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 188–195. <https://doi.org/10.1109/ROMAN.2018.8525769>
- Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06* (pp. 161–168). <https://doi.org/10.1145/1143844.1143865>

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, Y., Zhong, K., Zhang, J., Sun, Q., & Zhao, X. (2016). LSTM Networks for Mobile Human Activity Recognition. 50–53. <https://doi.org/10.2991/icaita-16.2016.13>
- Chinchor, N. (1992). MUC-4 evaluation metrics. *Proceedings of the 4th Conference on Message Understanding - MUC4 '92*, 22. <https://doi.org/10.3115/1072064.1072067>
- Choy, K. H. C., Yu, J., Hawkes, D., & Mayorov, D. N. (2012). Analysis of vigilant scanning behavior in mice using two-point digital video tracking. *Psychopharmacology*, 221(4), 649–657. <https://doi.org/10.1007/s00213-011-2609-5>
- Clary, D., & Kelly, D. M. (2016). Graded Mirror Self-Recognition by Clark's Nutcrackers. *Scientific Reports*, 6(1), 36459. <https://doi.org/10.1038/srep36459>
- Cobb, M. (2020). *The Idea of the Brain: A History*. Profile Books Ltd.
- Cotgreave, P., & Clayton, D. H. (1994). Comparative analysis of time spent grooming by birds in relation to parasite load. *Behavior*, 131, 171–187. <https://doi.org/10.1163/156853994X00424>
- Crispim Junior, C. F., Pederiva, C. N., Bose, R. C., Garcia, V. A., Lino-de-Oliveira, C., & Marino-Neto, J. (2012). ETHOW-ATCHER: Validation of a tool for behavioral and video-tracking analysis in laboratory animals. *Computers in Biology and Medicine*, 42(2), 257–264. <https://doi.org/10.1016/j.compbiomed.2011.12.002>
- Cruciani, F., Vafeiadis, A., Nugent, C., Cleland, I., McCullagh, P., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., & Hamzaoui, R. (2020). Feature learning for Human Activity Recognition using Convolutional Neural Networks. *CCF Transactions on Pervasive Computing and Interaction*, 2(1), 18–32. <https://doi.org/10.1007/s42486-020-00026-2>
- Dally, J. M., Emery, N. J., & Clayton, N. S. (2006). Food-Caching Western Scrub-Jays Keep Track of Who Was Watching When. *Science*, 312(5780), 1662–1665. <https://doi.org/10.1126/science.1126539>
- Dankert, H., Wang, L., Hoopfer, E. D., Anderson, D. J., & Perona, P. (2009). Automated Monitoring and Analysis of Social Behavior in *Drosophila*. *Nature Methods*, 6(4), 297–303. <https://doi.org/10.1038/nmeth.1310>
- Delius, J. D. (1967). Displacement activities and arousal. *Nature*, 214, 1259–1260. <https://doi.org/10.1038/2141259a0>
- Delius, J. D. (1988). Preening and associated comfort behavior in birds. *Ann. N. Y. Acad. Sci.*, 525, 40–55. <https://doi.org/10.1111/j.1749-6632.1988.tb38594.x>
- Delius, J. D., Perchard, R. J., & Emmerton, J. (1976). Polarized light discrimination by pigeons and an electroretinographic correlate. *Journal of Comparative and Physiological Psychology*, 90(6), 560–571. <https://doi.org/10.1037/h0077223>
- Dell, A. I., Bender, J. A., Branson, K., Couzin, I. D., de Polavieja, G. G., Noldus, L. P. J. J., Pérez-Escudero, A., Perona, P., Straw, A. D., Wikelski, M., & Brose, U. (2014). Automated image-based tracking and its application in ecology. *Trends in Ecology & Evolution*, 29(7), 417–428. <https://doi.org/10.1016/j.tree.2014.05.004>
- Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9_1
- Do, N.-T., Kim, S.-H., Yang, H.-J., & Lee, G.-S. (2020). Robust Hand Shape Features for Dynamic Hand Gesture Recognition Using Multi-Level Feature LSTM. *Applied Sciences*, 10(18), 6293. <https://doi.org/10.3390/app10186293>
- Du, Y., Mahdi, N., Paul, B., & Spetch, M. L. (2016). Cue salience influences the use of height cues in reorientation in pigeons (*Columba livia*). *Journal of Experimental Psychology. Animal Learning and Cognition*, 42(3), 273–280. <https://doi.org/10.1037/xan0000106>
- Evans, D. R., McArthur, S. L., Bailey, J. M., Church, J. S., & Reudink, M. W. (2015). A high-accuracy, time-saving method for extracting nest watch data from video recordings. *Journal of Ornithology*, 156(4), 1125–1129. <https://doi.org/10.1007/s10336-015-1267-5>
- Fagot, J., & Cook, R. G. (2006). Evidence for large long-term memory capacities in baboons and pigeons and its implications for learning and the evolution of cognition. *Proceedings of the National Academy of Sciences*, 103(46), 17564–17567. <https://doi.org/10.1073/pnas.0605184103>
- Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, 33(4), 917–963. <https://doi.org/10.1007/s10618-019-00619-1>
- Fawaz, H. I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2020). InceptionTime: Finding AlexNet for Time Series Classification. *Data Mining and Knowledge Discovery*, 34(6), 1936–1962. <https://doi.org/10.1007/s10618-020-00710-y>
- Fernández-Juricic, E., Siller, S., & Kacelnik, A. (2004). Flock density, social foraging, and scanning: an experiment with starlings. *Behav. Ecol.*, 15, 371–379. <https://doi.org/10.1093/behecol/arb017>
- Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1), 38–49. <https://doi.org/10.1017/S0376892997000088>
- Gallup, G. G. (1970). Chimpanzees: Self-Recognition. *Science*, 167(3914), 86–87. <https://doi.org/10.1126/science.167.3914.86>
- Gholamrezaei, M., & Taghi Almodarresi, S. M. (2019). Human Activity Recognition Using 2D Convolutional Neural Networks. *27th Iranian Conference on Electrical Engineering (ICEE)*, 1682–1686. <https://doi.org/10.1109/IranianCEE.2019.8786578>
- Gill, F. B., & Wolf, L. L. (1975). Economics of Feeding Territoriality in the Golden-Winged Sunbird. *Ecology*, 56(2), 333–345. <https://doi.org/10.2307/1934964>
- Gomez-Marin, A., Paton, J. J., Kampff, A. R., Costa, R. M., & Mainen, Z. F. (2014). Big behavioral data: Psychology, ethology and the foundations of neuroscience. *Nature Neuroscience*, 17(11), 1455–1462. <https://doi.org/10.1038/nn.3812>
- Green, J., Collins, C., Kyzar, E. J., Pham, M., Roth, A., Gaikwad, S., Cachat, J., Stewart, A. M., Landsman, S., Grieco, F., Tegelenbosch, R., Noldus, L. P. J. J., & Kalueff, A. V. (2012). Automated high-throughput neurophenotyping of zebrafish social behavior. *Journal of Neuroscience Methods*, 210(2), 266–271. <https://doi.org/10.1016/j.jneumeth.2012.07.017>
- Güntürkün, O., Koenen, C., Iovine, F., Garland, A., & Pusch, R. (2018). The neuroscience of perceptual categorization in pigeons: A mechanistic hypothesis. *Learning & Behavior*, 46(3), 229–241. <https://doi.org/10.3758/s13420-018-0321-6>
- Halvorsen, K., Johnston, C., Back, W., Stokes, V., & Lanshammar, H. (2008). Tracking the Motion of Hidden Segments Using Kinematic Constraints and Kalman Filtering. *Journal of Biomechanical Engineering*, 130(1). <https://doi.org/10.1115/1.2838035>
- Hirsch, J. (1986). Nothing in Neurobiology Makes Sense—Except in the Light of. *Behaviour*, 31(9), 674–676. <https://doi.org/10.1037/025029>
- Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1(1), 278–282. <https://doi.org/10.1109/ICDAR.1995.598994>
- Howard, J., & Gugger, S. (2020). fastai: A Layered API for Deep Learning. *Information*, 11(2), 108. <https://doi.org/10.3390/info11020108>

- Jethanandani, M., Perumal, T., Liaw, Y.-C., Chang, J.-R., Sharma, A., & Bao, Y. (2019). Binary Relevance Model for Activity Recognition in Home Environment using Ambient Sensors. *2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, 1–2. <https://doi.org/10.1109/ICCE-TW46550.2019.8991837>
- Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLOS Computational Biology*, *13*(1), e1005268. <https://doi.org/10.1371/journal.pcbi.1005268>
- Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S., & Branson, K. (2013). JAABA: Interactive machine learning for automatic annotation of animal behavior. *Nature Methods*, *10*(1), 64–67. <https://doi.org/10.1038/nmeth.2281>
- Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45. <https://doi.org/10.1115/1.3662552>
- Kilian, A., Yaman, S., von Fersen, L., & Güntürkün, O. (2003). A bottlenose dolphin discriminates visual stimuli differing in numerosity. *Animal Learning & Behavior*, *31*(2), 133–142. <https://doi.org/10.3758/BF03195976>
- Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, *26*(9), 1011–1013. <https://doi.org/10.1038/nbt0908-1011>
- Koch, P., Dreier, M., Maass, M., Bohme, M., Phan, H., & Mertins, A. (2019). A Recurrent Neural Network for Hand Gesture Recognition based on Accelerometer Data. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5088–5091. <https://doi.org/10.1109/EMBC.2019.8856844>
- Kolekar, M. H., & Dash, D. P. (2016). Hidden Markov Model based human activity recognition using shape and optical flow based features. *2016 IEEE Region 10 Conference (TENCON)*, 393–397. <https://doi.org/10.1109/TENCON.2016.7848028>
- Kraft, F.-L., Forštová, T., Utku Urhan, A., Exnerová, A., & Brodin, A. (2017). No evidence for self-recognition in a small passerine, the great tit (*Parus major*) judged from the mark/mirror test. *Animal Cognition*, *20*(6), 1049–1057. <https://doi.org/10.1007/s10071-017-1121-7>
- Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490. <https://doi.org/10.1016/j.neuron.2016.12.041>
- Kralj-Fiser, S., Scheiber, I. B. R., Kotrschal, K., Weiss, B. M., & Wascher, C. A. F. (2010). Glucocorticoids enhance and suppress heart rate and behaviour in time dependent manner in greylag geese (*Anser anser*). *Physiol. Behav.*, *100*, 394–400. <https://doi.org/10.1016/j.physbeh.2010.04.005>
- Kravitz, A. V., Freeze, B. S., Parker, P. R. L., Kay, K., Thwin, M. T., Deisseroth, K., & Kreitzer, A. C. (2010). Regulation of parkinsonian motor behaviors by optogenetic control of basal ganglia circuitry. *Nature*, *466*(7306), 622–626. <https://doi.org/10.1038/nature09159>
- Lee, S.-M., Yoon, S. M., & Cho, H. (2017). Human activity recognition from accelerometer data using Convolutional Neural Network. *IEEE International Conference on Big Data and Smart Computing (BigComp)*, 131–134. <https://doi.org/10.1109/BIGCOMP.2017.7881728>
- Lindburg, D. G. (1969). Behavior of infant rhesus monkeys with thalidomide-induced malformations: A pilot study. *Psychonomic Science*, *15*(1), 55–56. <https://doi.org/10.3758/BF03336196>
- Lochmatter, T., Roduit, P., Cianci, C., Correll, N., Jacot, J., & Martinoli, A. (2008). SwisTrack—A Flexible Open Source Tracking Software for Multi-Agent Systems. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4004–4010. <https://doi.org/10.1109/IROS.2008.4650937>
- Luxem, K., Fuhrmann, F., Kürsch, J., Remy, S., & Bauer, P. (2020). Identifying Behavioral Structure from Deep Variational Embeddings of Animal Motion. *bioRxiv*. <https://doi.org/10.1101/2020.05.14.095430>
- Luyten, L., Schroyens, N., Hermans, D., & Beckers, T. (2014). Parameter optimization for automated behavior assessment: Plug-and-play or trial-and-error? *Frontiers in Behavioral Neuroscience*, *8*. <https://doi.org/10.3389/fnbeh.2014.00028>
- Mannini, A., & Sabatini, A. M. (2010). Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers. *Sensors*, *10*(2), 1154–1175. <https://doi.org/10.3390/s100201154>
- Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281–1289. <https://doi.org/10.1038/s41593-018-0209-y>
- Mazur-Milecka, M., & Ruminski, J. (2017). Automatic analysis of the aggressive behavior of laboratory animals using thermal video processing. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. *IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2017, 3827–3830. <https://doi.org/10.1109/EMBC.2017.8037691>
- Mehmood, A., Iqbal, M., Mehmood, Z., Irtaza, A., Nawaz, M., Nazir, T., & Masood, M. (2021). Prediction of Heart Disease Using Deep Convolutional Neural Networks. *Arabian Journal for Science and Engineering*, *46*(4), 3409–3422. <https://doi.org/10.1007/s13369-020-05105-1>
- Mekruksavanich, S., & Jitpattanukul, A. (2021). Deep Convolutional Neural Network with RNNs for Complex Activity Recognition Using Wrist-Worn Wearable Sensor Data. *Electronics*, *10*(14), 1685. <https://doi.org/10.3390/electronics10141685>
- Miller, E. H. (1988). *Description of Bird Behavior for Comparative Purposes*. In R. F. Johnston (Ed.), *Current Ornithology* (pp. 347–394). Springer. https://doi.org/10.1007/978-1-4615-6787-5_9
- Miller, N., & Gerlai, R. (2012). Automated Tracking of Zebrafish Shoals and the Analysis of Shoaling Behavior. In A. V. Kalueff & A. M. Stewart (Eds.), *Zebrafish Protocols for Neurobehavioral Research* (pp. 217–230). Humana Press. https://doi.org/10.1007/978-1-61779-597-8_16
- Miri, A., Warriner, C. L., Seely, J. S., Elsayed, G. F., Cunningham, J. P., Churchland, M. M., & Jessell, T. M. (2017). Behaviorally Selective Engagement of Short-Latency Effector Pathways by Motor Cortex. *Neuron*, *95*(3), 683–696.e11. <https://doi.org/10.1016/j.neuron.2017.06.042>
- Moyer, B. R., Rock, A. N., & Clayton, D. H. (2003). Experimental Test of the Importance of Preen Oil in Rock Doves (*Columba livia*). *The Auk*, *120*(2), 490–496. <https://doi.org/10.1093/auk/120.2.490>
- Mundorf, A., Matsui, H., Ocklenburg, S., & Freund, N. (2020). Asymmetry of turning behavior in rats is modulated by early life stress. *Behavioral Brain Research*, *393*, 112807. <https://doi.org/10.1016/j.bbr.2020.112807>
- Murad, A., & Pyun, J.-Y. (2017). Deep Recurrent Neural Networks for Human Activity Recognition. *Sensors*, *17*(11), 2556. <https://doi.org/10.3390/s17112556>
- Murakami, K., & Taguchi, H. (1991). Gesture recognition using recurrent neural networks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 237–242. <https://doi.org/10.1145/108844.108900>
- Naqvi, R. A., Arsalan, M., Rehman, A., Rehman, A. U., Loh, W.-K., & Paul, A. (2020). Deep Learning-Based Drivers Emotion Classification System in Time Series Data for Remote Applications. *Remote Sensing*, *12*(3), 587. <https://doi.org/10.3390/rs12030587>
- Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., & Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation

- across species and behaviors. *Nat. Protoc.*, 14, 2152–2176. <https://doi.org/10.1038/s41596-019-0176-0>
- Nema, S., Hasan, W., Bhargava, A., & Bhargava, Y. (2016). A novel method for automated tracking and quantification of adult zebrafish behavior during anxiety. *Journal of Neuroscience Methods*, 271, 65–75. <https://doi.org/10.1016/j.jneumeth.2016.07.004>
- Nematallah, H., Rajan, S., & Cretu, A.-M. (2019). Logistic Model Tree for Human Activity Recognition Using Smartphone-Based Inertial Sensors. *IEEE SENSORS, 2019*, 1–4. <https://doi.org/10.1109/SENSORS43011.2019.8956951>
- Nilsson, S. R., Goodwin, N. L., Choong, J. J., Hwang, S., Wright, H. R., Norville, Z. C., Tong, X., Lin, D., Bentzley, B. S., Eshel, N., McLaughlin, R. J., & Golden, S. A. (2020). Simple Behavioral Analysis (SimBA) – an open source toolkit for computer classification of complex social behaviors in experimental animals. *BioRxiv*, 2020(04), 19.049452. <https://doi.org/10.1101/2020.04.19.049452>
- Noldus, L. P. J. J., Spink, A. J., & Tegelenbosch, R. A. J. (2001). EthoVision: A versatile video tracking system for automation of behavioral experiments. *Behavior Research Methods, Instruments, & Computers*, 33(3), 398–414. <https://doi.org/10.3758/BF03195394>
- Oguiza, I. (2020). tsai—A state-of-the-art deep learning library for time series and sequential data. github.com/timeseriesAI/tsai
- Packheiser, J., Güntürkün, O., & Pusch, R. (2019). Renewal of extinguished behavior in pigeons (*Columba livia*) does not require memory consolidation of acquisition or extinction in a free-operant appetitive conditioning paradigm. *Behavioural Brain Research*, 370, 111947. <https://doi.org/10.1016/j.bbr.2019.111947>
- Packheiser, J., Donoso, J. R., Cheng, S., Güntürkün, O., & Pusch, R. (2021). Trial-by-trial dynamics of reward prediction error-associated signals during extinction learning and renewal. *Progress in Neurobiology*, 197, 101901. <https://doi.org/10.1016/j.pneurobio.2020.101901>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* 32 (pp. 8024–8035).
- Paul, P., & George, T. (2015). An effective approach for human activity recognition on smartphone. *IEEE International Conference on Engineering and Technology (ICETECH)*, 1–3. <https://doi.org/10.1109/ICETECH.2015.7275024>
- Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex* (pp. xv, 430). Oxford Univ. Press.
- Pearce, J. M., Esber, G. R., George, D. N., & Haselgrove, M. (2008). The nature of discrimination learning in pigeons. *Learning & Behavior*, 36(3), 188–199. <https://doi.org/10.3758/LB.36.3.188>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830.
- Peebles, D., & Cooper, R. P. (2015). Thirty Years After Marr's Vision: Levels of Analysis in Cognitive Science. *Topics in Cognitive Science*, 7(2), 187–190. <https://doi.org/10.1111/tops.12137>
- Peng, M., Zhang, C., Dong, Y., Zhang, Y., Nakazawa, H., Kaneki, M., Zheng, H., Shen, Y., Marcantonio, E. R., & Xie, Z. (2016). Battery of behavioral tests in mice to study postoperative delirium. *Scientific Reports*, 6, 29874. <https://doi.org/10.1038/srep29874>
- Pepperberg, I. M., Garcia, S. E., Jackson, E. C., & Marconi, S. (1995). Mirror use by African Grey parrots (*Psittacus erithacus*). *Journal of Comparative Psychology*, 109(2), 182–195. <https://doi.org/10.1037/0735-7036.109.2.182>
- Pérez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S., & de Polavieja, G. G. (2014). idTracker: Tracking individuals in a group by automatic identification of unmarked animals. *Nature Methods*, 11(7), 743–748. <https://doi.org/10.1038/nmeth.2994>
- Peterson, A. H., & Martinez, T. (2005). Estimating The Potential for Combining Learning Models. In *Proceedings of the ICML Workshop on Meta-Learning*, pages, 68–75, 2005.
- Piato, Á. L., Capiotti, K. M., Tamborski, A. R., Osés, J. P., Barcellos, L. J. G., Bogo, M. R., Lara, D. R., Vianna, M. R., & Bonan, C. D. (2011). Unpredictable chronic stress model in zebrafish (*Danio rerio*): Behavioral and physiological responses. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 35(2), 561–567. <https://doi.org/10.1016/j.pnpb.2010.12.018>
- Pienaar, S. W., & Malekian, R. (2019). Human Activity Recognition Using LSTM-RNN Deep Neural Network Architecture. *ArXiv:1905.00599* [Cs, Eess, Stat].
- Piryonesi, S. M., & El-Diraby, T. E. (2020). Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation Engineering, Part B: Pavements*, 146(2), 04020022. <https://doi.org/10.1061/JPEODX.0000175>
- Pollok, B., Prior, H., & Güntürkün, O. (2000). Development of object permanence in food-storing magpies (*Pica pica*). *Journal of Comparative Psychology*, 114(2), 148–157. <https://doi.org/10.1037/0735-7036.114.2.148>
- Prior, H., Schwarz, A., & Güntürkün, O. (2008). Mirror-Induced Behavior in the Magpie (*Pica pica*): Evidence of Self-Recognition. *PLOS Biology*, 6(8), e202. <https://doi.org/10.1371/journal.pbio.0060202>
- Reiss, D., & Marino, L. (2001). Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence. *Proceedings of the National Academy of Sciences*, 98(10), 5937–5942. <https://doi.org/10.1073/pnas.101086398>
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, 28.
- Rijsbergen, V. C. J. (1979). *Information retrieval*. Butterworths.
- Rodríguez, A., Ortega-álvaro, A., Sola, R., Micó, J. A., & Trelles, O. (2004). Automatic tracking analysis in Morris water maze biomedical videos. In *Proc. of the International Conference on Visualization, Imaging and Image Processing*.
- Rose, J., Schmidt, R., Grabemann, M., & Güntürkün, O. (2009). Theory meets pigeons: The influence of reward-magnitude on discrimination-learning. *Behavioural Brain Research*, 198(1), 125–129. <https://doi.org/10.1016/j.bbr.2008.10.038>
- Scarf, D., Boy, K., Reinert, A. U., Devine, J., Güntürkün, O., & Colombo, M. (2016). Orthographic processing in pigeons (*Columba livia*). *Proceedings of the National Academy of Sciences*, 113(40), 11272–11276. <https://doi.org/10.1073/pnas.1607870113>
- Singh, D., Merdivan, E., Psychoula, I., Kropf, J., Hanke, S., Geist, M., & Holzinger, A. (2017). Human Activity Recognition using Recurrent Neural Networks. *ArXiv:1804.07144* [Cs, Stat], 10410, 267–274. https://doi.org/10.1007/978-3-319-66808-6_18
- Singh, S., Kaur, H., & Sandhir, R. (2016). Fractal dimensions: A new paradigm to assess spatial memory and learning using Morris water maze. *Behavioural Brain Research*, 299, 141–146. <https://doi.org/10.1016/j.bbr.2015.11.023>

- Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis* (p. 457). Appleton-Century.
- Smith, L. N. (2017). Cyclical Learning Rates for Training Neural Networks. *ArXiv:1506.01186* [Cs].
- Smith, L. N. (2018). A disciplined approach to neural network hyperparameters: Part 1 -- learning rate, batch size, momentum, and weight decay. *ArXiv:1803.09820* [Cs, Stat].
- Smith, L. N., & Topin, N. (2018). *Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates*. openreview.net/forum?id=H1A5ztj3b
- Sridhar, V. H., Roche, D. G., & Gingins, S. (2019). Tracktor: Image-based automated tracking of animal movement and behavior. *Methods in Ecology and Evolution*, 10(6), 815–820. <https://doi.org/10.1111/2041-210X.13166>
- Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04), 687–719. <https://doi.org/10.1142/S0218001409007326>
- Thorndike, E. L. (1898). Animal Intelligence: An Experimental Study of the Associative Processes in Animals. *Psychological Review*, 5(5), 551–553. <https://doi.org/10.1037/h0067373>
- Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift Für Tierpsychologie*, 20(4), 410–433. <https://doi.org/10.1111/j.1439-0310.1963.tb01161.x>
- Tosi, M. V., Ferrante, V., Mattiello, S., Canali, E., & Verga, M. (2006). Comparison of video and direct observation methods for measuring oral behaviour in veal calves. *Italian Journal of Animal Science*, 5(1), 19–27. <https://doi.org/10.4081/ijas.2006.19>
- Troje, N. F., Huber, L., Loidolt, M., Aust, U., & Fieder, M. (1999). Categorical learning in pigeons: The role of texture and shape in complex static stimuli. *Vision Research*, 39(2), 353–366. [https://doi.org/10.1016/S0042-6989\(98\)00153-9](https://doi.org/10.1016/S0042-6989(98)00153-9)
- Vallortigara, G., Regolin, L., & Marconato, F. (2005). Visually Inexperienced Chicks Exhibit Spontaneous Preference for Biological Motion Patterns. *PLOS Biology*, 3(7), e208. <https://doi.org/10.1371/journal.pbio.0030208>
- Vaughan, W., & Greene, S. L. (1984). Pigeon visual memory capacity. *Journal of Experimental Psychology: Animal Behavior Processes*, 10(2), 256–271. <https://doi.org/10.1037/0097-7403.10.2.256>
- von Frisch, K. (1967). *The dance language and orientation of bees*. Belknap Press of Harvard University Press.
- Walter, T., & Couzin, I. D. (2020). TRex, a fast multi-animal tracking system with markerless identification, and 2D estimation of posture and visual fields. *BioRxiv*, 2020(10), 14.338996. <https://doi.org/10.1101/2020.10.14.338996>
- Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119, 3–11. <https://doi.org/10.1016/j.patrec.2018.02.010>
- Wilzeck, C., Wiltschko, W., Güntürkün, O., Buschmann, J.-U., Wiltschko, R., & Prior, H. (2010). Learning of magnetic compass directions in pigeons. *Animal Cognition*, 13(3), 443–451. <https://doi.org/10.1007/s10071-009-0294-0>
- Wittek, N., Matsui, H., Kessel, N., Oeksuez, F., Güntürkün, O., & Anselme, P. (2021). Mirror Self-Recognition in Pigeons: Beyond the Pass-or-Fail Criterion. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.669039>
- Wu, S.-T., & Marquez, M. R. G. (2003). A non-self-intersection Douglas-Peucker algorithm. *16th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2003)*, 60–66. <https://doi.org/10.1109/SIBGRA.2003.1240992>
- Xu, L., Yang, W., Cao, Y., & Li, Q. (2017). Human activity recognition based on random forests. *13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 548–553. <https://doi.org/10.1109/FSKD.2017.8393329>
- Yamanaka, O., & Takeuchi, R. (2018). UMATracker: An intuitive image-based tracking platform. *The Journal of Experimental Biology*, 221(Pt, 16). <https://doi.org/10.1242/jeb.182469>
- Zentall, T. R., Galizio, M., & Critchfield, T. S. (2013). Categorization, Concept Learning, and Behavior Analysis: An Introduction. *Journal of the Experimental Analysis of Behavior*, 78(3), 237–248. <https://doi.org/10.1901/jeab.2002.78-237>

Open Practices Statement

All data have been made publicly available at OSF and can be accessed at <https://osf.io/4285v/>. In addition, the corresponding code is available at <https://github.com/kiview/winkie>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.