BOVIDS: A deep learning-based software for pose estimation to evaluate nightly behavior and its application to Common Elands (*Tragelaphus oryx*) in zoos

Jennifer Gübert[1*], Max Hahn-Klimroth[2], Paul W. Dierkes[1]

[1]Faculty of Biological Sciences, Bioscience Education and Zoo Biology, Goethe University, Frankfurt, Germany

[2]Faculty of Computer Science, TU Dortmund University, Dortmund, Germany

*** Correspondence:**
Jennifer Gübert
guebert@bio.uni-frankfurt.de

Keywords: *Tragelaphus oryx*, nightly behavior, REM sleep, posture estimation, deep learning, video action classification

BOVIDS: A deep learning-based software for pose estimation to evaluate nightly behavior and its application to Common Elands (*Tragelaphus oryx*) in zoos

Abstract

Only a few studies on the nocturnal behavior of African ungulates exist so far, with mostly small sample sizes. For a comprehensive understanding of nocturnal behavior, this database needs to be expanded. Zoo animals offer a good opportunity to lay the corresponding foundations. The results can provide clues for the study of wild animals and furthermore contribute to a better understanding of animal welfare and better husbandry conditions in zoos. To tackle this open question, we developed a stand-alone open-source software based on deep learning techniques, named BOVIDS (**B**ehavioral **O**bservations by **V**ideos and **I**mages using a **D**eep-Learning **S**oftware). This software is used to identify ungulates in their enclosure and to determine crucial behavioral poses on video material with an accuracy of 99.4%. A case study on 25 Common Elands (*Tragelaphus oryx*) out of 5 EAZA zoos with a total of 11,411 hours video material out of 822 nights is conducted, yielding the first detailed description of the nightly behavior of Common Elands. Our results indicate that age and sex are influencing factors on the nocturnal activity budget, the length of behavioral phases as well as the number of phases per behavioral state during the night. Finally, the results suggest the existence of species-specific rhythms that open future research directions.

1 Introduction

The nocturnal behavior of many African mammals is poorly studied. It is known that the behavioral patterns can vary greatly between day and night, as many large herbivorous mammals spend especially in winter most of their sleeping time during night, while the activity patterns emerge primarily at daytime (Bennie et al., 2014; Gravett et al., 2017; Davimes et al., 2018; Wu et al., 2018). For a comprehensive understanding of diurnal rhythms, a behavioral description of the entire diurnal cycle is necessary. So far, especially the nocturnal behavior is little studied, not only in the free-range but also in zoos. A major advantage of observing zoo animals at night rather than animals in their natural habitat (Ryder and Feistner, 1995) is that it is much easier to install observation equipment. In order not to disturb the animals, camera recordings are a good means of data collection in this case. Thus, zoos provide a good basis for describing the animals' nocturnal behavior and the results can subsequently serve as starting information for observations in the field (Burger et al., 2020). In addition, a deeper knowledge of nocturnal behavior could contribute information to further improve animal management and husbandry in zoos (Brando and Buchanan-Smith, 2018) and provide conclusions on animal welfare (Walsh et al., 2019). For example, REM sleep appears to be an important indicator of stress in giraffes (Sicks, 2016), which can be measured by non-invasive methods.

To describe nocturnal behavior unambiguously, a lot of data is needed, especially because there are few comparisons in literature. Furthermore, it would not only be useful to examine a lot of individuals from one species to compensate for the lack of comparable data, but also many nights of every individual would have to be analyzed to accurately describe the average behavior. Additionally, it is necessary to obtain data not only on one but many different species to close the existing knowledge gap. However, the extraction of meaningful information as well as a detailed evaluation of a mass of recorded data requires modern techniques to automate parts of this data mining process (Norouzzadeh et al., 2018; Lürig et al., 2021). Fortunately, in the last decade, various computer vision and deep learning techniques found their way into behavioral biology and ecology (Dell et al., 2014; Valletta et al., 2017; Eikelboom et al., 2019; Chakravarty et al., 2020; Gerovichev et al., 2021; Norouzzadeh et al., 2021), providing amazing results and facilitating the task of dealing with a large dataset dramatically. Unfortunately, automatization of the evaluation of video recordings is challenging if the video recordings suffer from a low framerate, much background noise or heavy truncation effects, as is usual in observations in stables as zoo enclosures, or even in free-range installments. Therefore, only a few of those computer systems are applicable for the challenging data generated by field studies or records in a variety of zoo enclosures.

One of the two main objectives of this work tackles this challenge by making BOVIDS (**B**ehavioral **O**bservations by **V**ideos and **I**mages using a **D**eep-Learning **S**oftware), a stand-alone software based on deep learning techniques, available. To the best of our knowledge, this is the first fully open-source software tackling the task of evaluating the nightly behavior of stalled animals that contains functionalities required for data preparation, training of the deep learning parts, data prediction and data presentation. More precisely, BOVIDS can be used to evaluate video recordings of stalled ungulates recorded at 1 fps regarding two classification tasks: "binary classification" and "total classification" (Hahn-Klimroth et al., 2021). In the total classification task, BOVIDS predicts one out of the following poses per seven seconds of video: Standing, Lying - head up (LHU), Lying - head down (LHD), being out of view (Out). The binary classification task asks only for one label of Standing, Lying (LHU and LHD) or being out of view (Out) and is useful to study rhythms. The software can be divided into four components:

- BOV 1. Data collection,
- BOV 2. Object detection (OD),
- BOV 3. Action classification (AC),
- BOV 4. Data prediction.

While one part of BOV 4 is a significantly improved and extended version of work presented in an earlier contribution (Hahn-Klimroth et al., 2021), the newly developed components BOV 1 - BOV 3 allow an interested user to apply the complete deep learning prediction system comfortably to their own data. All discussed software as well as detailed instructions can be found in our GitHub repository: https://github.com/Klimroth/BOVIDS.

This paper not only extends and improves the previous software but explains how BOVIDS can be applied by behavioral biologists to their own data. To this end, BOVIDS is applied to data of Common Elands (*Tragelaphus oryx*) showing the power of the obtained method.

More precisely, a case study on the nocturnal activity budget of Common Elands is the second main objective of the present work and has a dual purpose. First, in the case study over 11.000 hours (822 nights) of video material from five different EAZA zoos were evaluated, a task that seems inaccessible in the absence of automatic evaluation. Second, it shows how BOVIDS can be used to observe and analyze several important behavioral biological key figures of nocturnal activity. Finally, and at least as importantly, to the best of our knowledge, the case study provides the first excessive and detailed description of important aspects of the nocturnal behavior of Common Elands. This description contains activity budgets, a visualization of the Standing-Lying rhythm as well as an analysis of the possible influencing factors age, sex, and zoo husbandry.

As mentioned earlier, several computational systems have found their way into behavioral biology and ecology (Dell et al., 2014; Valletta et al., 2017; Eikelboom et al., 2019; Chakravarty et al., 2020; Norouzzadeh et al., 2021). Such systems are explicitly designed with respect to the underlying data. In the easiest tasks, cameras can be installed in a laboratory such that the recordings feature a high contrast between animals and the background as well as other laboratory conditions like a given steady camera angle and low background noise. Examples for such systems working with data of *Drosophila*-flies or mice are *JAABA* (Kabra et al., 2013), *DeepBehavior* (Graving et al., 2019) and *SLEAP* (Pereira et al., 2020). When data is recorded either in the natural habitat or in different zoo enclosures, it is much more challenging due to variations in weather, brightness, and background. Not to forget, different cameras can rarely be adjusted in a way such that the recording angle matches given requirements or to ensure that animals are not highly truncated. One approach under varying brightness conditions distinguishes the poses "lying" and "standing" of cows in free-stall stables (Porto et al., 2013). Finally, one of the most impressive success stories might be the work by Norouzzadeh et al. (2018; 2021) whose system is able to automatically detect and count different species and some shown behaviors using camera trap images of the Serengeti dataset (Swanson et al., 2015).

2 Material and Methods

In the first section *Data evaluation* methods and material used to collect the data and to evaluate the findings statistically are presented. Subsequently, the behavioral states of interest are defined properly in section *Ethogram* whereas section *BOVIDS* introduces and describes the software package BOVIDS which is the main technical contribution of the present paper.

2.1 Data evaluation

The dataset includes nights of 25 Common Elands (*Tragelaphus oryx*) whereas the number of nights per individual ranges from 15 to 49. In total, 822 nights with 11,411 hours of video material are present. The data was collected in winter seasons between 2017 and 2020 in a total of five EAZA zoos in Germany (Allwetterzoo Münster, Erlebnis-Zoo Hannover, Opel-Zoo Kronberg, Zoo Dortmund and Zoom Erlebniswelt Gelsenkirchen). A detailed overview about the used data is given in Table 1 in the appendix. For further analysis the individuals are categorized as follows: 'young', ranging from birth until the time of weaning with about six months, 'subadult', older than six months until sexual maturity with about two years of age and 'adult' afterwards. Those categories are chosen accordingly to information distributed across multiple prior works (Puschmann et al., 2009; Groves and Leslie Jr, 2011; Tacutu et al., 2013; Myers et al., 2021).

All collected data is in the form of video recordings. The cameras used are capable of night vision due to built-in infrared emitters (Lupus LE139HD or Lupus LE338HD with the recording device LUPUSTEC LE800HD or TECHNAXX PRO HD 720P). The recordings are made with a frame rate of 1 fps and the resolution ranges from 704x576 px to 1920x1080 px. Recording takes place in the stable during night, the time of the absence of animal keepers, which mostly ranges from 17:00 to 07:00 (14 hours). In some cases, the recording time is 18:00 to 07:00 (13 hours).

The data was recorded continuously providing an exact time span for every behavior with a start and an end time (Martin and Bateson, 2015). The manually annotation was governed by the open-source program BORIS, Version 7.7.3 (Friard and Gamba, 2016) and consists of 2,374 hours of video material out of 170 nights. BOVIDS requires the use of multiple deep neural networks for object detection (OD) and action classification (AC) as explained in Hahn-Klimroth et al. (2021) and in the following section. To train an initial object detection network, at least 400 images of every enclosure were annotated using LabelImg (Tzutalin, 2015) resulting in 11,326 images of Common Elands and 49,437 images of various African ungulates as already elaborated by Hahn-Klimroth et al. (2021). Following the prescribed approach, the initial action classification networks were not only trained using 170 recordings (66,466 images) of Common Elands but also 113,407 images of other African ungulates with comparable postures. Furthermore, two rounds of offline hard example mining (OHEM) were conducted using additionally 14,381 images of Common Elands and 50,262 images of other African ungulates. Finally, the action classifiers used for Common Elands stalled together were fine-tuned by

24,304 images stemming from manually annotated video files and 7,377 images generated through OHEM. Detailed information can be found in Table 1 in the appendix.

All statistical analysis is conducted with the software R Studio (R Core Team, 2014) and the figures, which are not given by BOVIDS, are produced using the core functionalities of R and the package ggplot2 (Wickham, 2016). Statistical tests are performed differently for continuous and ordinal data. To conduct a two-factor analysis of variance (ANOVA) on continuous data, normality is required which is tested by Shapiro-Wilk test for any behavior class. In case of significant deviation from normality ($p < 0.05$), a normality transformation is applied to the data by R's "bestNormalize" package (Peterson and Cavanaugh, 2020). To analyze differences between multiple groups on ordinal data, a Kruskal-Wallis test is applied. Finally, as post-hoc tests on all pairs of potentially significant factors, a collection of unpaired t-tests is applied in the continuous case and, respectively, a collection of Wilcoxon tests in the ordinal case. The alpha level is adjusted by the Bonferroni-Holm adjustment in each case.

2.2 Ethogram

The focus of this paper is to distinguish between four postures: Standing, Lying – head up (LHU), Lying – head down (LHD) and out of view (Out). The last category is used if the animal is not present in the stable and should also be used if only a small part of the animal is visible, from which the behavior cannot be determined. Furthermore, Lying in the binary classification system is defined as the union of LHU and LHD. The binary classification is helpful to analyze rhythms over the night as the categories "activity" and "rest" are the most prominently measured behavior stages to examine diurnal rhythms (Merrow et al., 2005). In the following ethogram, based on that of Hahn-Klimroth et al. (2021), the three behavioral states are defined and shown in Figure 1.

Standing: The animal stands in an upright position on all four hooves. The exact behavior is neglected, thus the animal could be, for instance, feeding, resting, or ruminating.

Lying – head up (LHU): The animal lies down, and its head is lifted. The behavioral state does not distinguish if the animal is awake or in non-REM sleep. As before, the precise behavior is neglected.

Lying – head down (LHD): The animal is lying with its head resting on the ground. The head's position is beside the body or sometimes in front of it.

It is crucial to notice that LHD is the typical REM (rapid eye movement) sleep posture. REM sleep is recognized through various behavioral components as the animal is lying with its head resting due to postural atonia (Lima et al., 2005; Zepelin et al., 2005). This characteristically REM sleep position can be used to estimate the REM sleep, a common approach in the study of behavior of Common Eland's (Zizkova et al., 2013) and cows (Ternman et al., 2014).

Figure 1. The three observed behavioral states: Standing, Lying - head up, Lying - head down, from left to right of Common Elands.

2.3 BOVIDS

BOVIDS is an end-to-end software package which automatically detects the poses of interest in videos. The detection itself is based on a combination of two deep-learning steps (object detection and action classification) governed by state-of-the-art deep neural networks. In the following, the single parts of BOVIDS will be introduced. To this end, it will be first described what the goal and functionality of BOVIDS are.

2.3.1 Overview

BOVIDS is used to automatically annotate the behavior of ungulates on video recordings. Those recordings are required to come as appropriately structured and formatted video files and BOV 1 contains python scripts that can generate the necessary video files out of the recordings by the LUPUS observation system. To annotate new data automatically, the prediction pipeline of BOVIDS (BOV 4) uses a composition of two stages of deep neural networks, an object detector to find individuals on the frames of the videos and action classifiers that are responsible for the posture estimation. The prediction pipeline itself will be discussed in detail later. Before being able to use the prediction pipeline, those deep neural networks need to be trained on manually annotated data. BOV 2 contains the necessary ingredients to train an object detector based on a recent yolov4 implementation (Taipingeric, 2020), while BOV 3 provides the software required to obtain decent action classifiers which are EfficientNet-B3 CNNs (Tan and Le, 2019). Beside the necessary scripts for training the networks, BOVIDS contains various tools to generate the training sets, organize the data and fine-tune networks like, for instance, by offline hard example mining (Felzenszwalb et al., 2010). Finally, multiple tools to measure the accuracy of the prediction and to detect systematic errors by BOVIDS are provided as well as tools to represent and visualize the data that are a good starting point to apply further statistical methods (BOV 4). Subsequently, components BOV 1 - BOV 4 are described in detail and a description on how to successfully apply BOVIDS to new data is given. A visualization of the complete process is given in Figure 2.

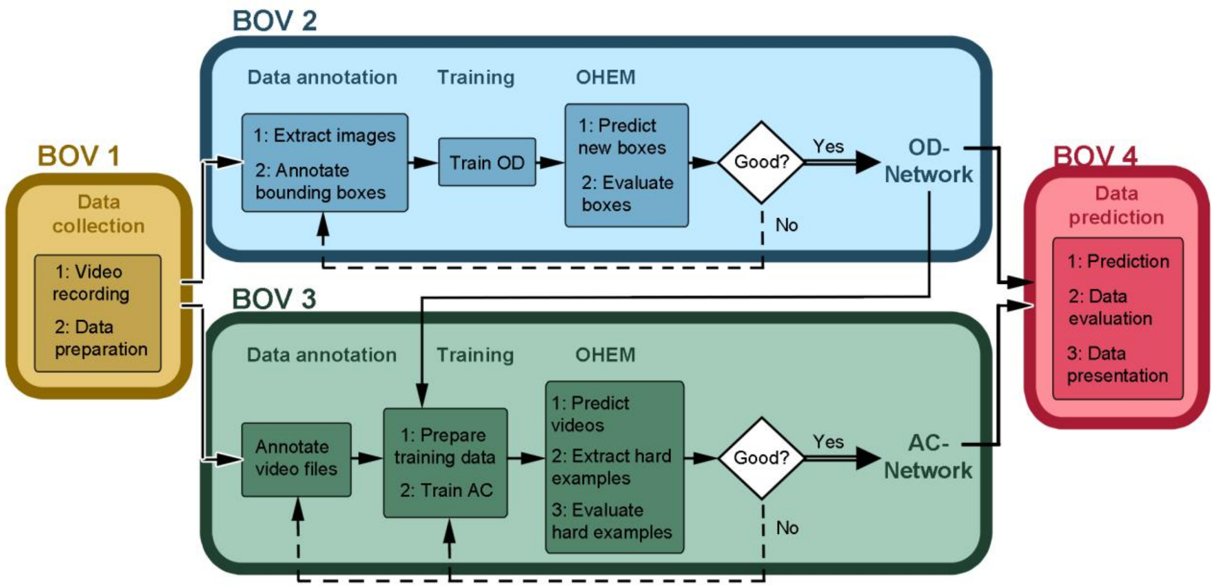

Figure 2. Overview of the System BOVIDS and all its categories.

2.3.2 BOV 1: Data preparation

This step creates a collection of video files, one per night. If a user records the data by the LUPUS observation system, BOVIDS provides a python script that can concatenate and convert the output of LUPUS into a collection of avi-files. If some data is missing, the missing frames can be filled with a sequence of black frames to ensure a joint observation time over all video files. Such sequences of black frames will be labeled as Out by BOVIDS during prediction and therefore represent reality quite well.

2.3.3 BOV 2: Object detection (OD)

The final object detector is trained in multiple steps which will be first mentioned shortly and afterwards described in detail. The procedure goes along as follows:

- OD 1. Manual annotation of images.
- OD 2. Train a first version of the object detector.
- OD 3. Offline hard example mining (OHEM).
  - a. Automatic annotation of unseen data.
  - b. Evaluation of the suggested bounding boxes.
  - c. Retrain the deep neural network.

In the initial annotation task (OD 1), between 400 and 800 images per enclosure should be sampled stemming from four to six videos over the observation period to increase the data variability. The number of images sampled in total depends on how much data there is overall, how difficult the detection appears to be and whether individuals need to be distinguished. Those images are now annotated manually by a freely available third-party software called LabelImg (Tzutalin, 2015), and after a few data preparation steps, the initial training can be performed (OD 2).

To run the so-called "offline hard example mining" (Felzenszwalb et al., 2010), in short OHEM (OD 3), the freshly trained object detector is used to automatically annotate 300 - 600 images out of unseen videos of the same set of enclosures (OD 3a). The quality of each such automatically drawn bounding box is evaluated manually by assigning one out of four classes (good, okay, poor, swapped) visualized in Figure 3 (OD 3b). If the bounding boxes are, overall, satisfyingly accurate, the procedure stops at this point. Otherwise, the poorly evaluated bounding boxes are corrected manually using LabelImg again. Those bounding boxes can be seen as "hard examples" as the current version of the object detector struggles at prediction. The freshly corrected annotations together with the well evaluated bounding boxes are used to increase the training set of the object detector and the object detector is trained on this new, extended set. This procedure can, in principle, be repeated until satisfying results are achieved but our experience shows that, once the used object detector generalizes decently, one round should be enough to achieve a sufficient accuracy. After having trained an accurately working object detector, this object detector is one ingredient required to train the action classifiers.

Figure 3. Example of the four classes that can be given in evaluation, good (green), okay (yellow), bad (red) and swapped (blue).

2.3.4 BOV 3: Action classification (AC)

The action classifier's goal is to predict the pose of an individual on a single image (single-frame, SF), respectively on four subsequent images placed next to each other (multiple-frame, MF). To achieve a well performing action classifier, the procedure reads as follows:

AC 1. Annotation of video files.

AC 2. Training of a first version of the ACs.

a. Preparation of an initial training set.

b. Training of the ACs.

AC 3. One or multiple rounds of OHEM

a. Prediction of many new video files.

b. Extracting hard as well as random examples as single images.

c. Manually evaluating the performance on those examples.

d. Retrain the network based on the evaluated images.

When starting from scratch, it is most convenient to annotate the behavior of each single frame of a video by annotating the whole video (AC 1), for instance using the third-party software BORIS (Friard and Gamba, 2016). After this initial annotation, the output of BORIS needs to be converted into a machine-readable labeling of each time-interval of a specific video. Then, equally many images representing the time-intervals of each posture (Standing, LHU, LHD) are extracted automatically from the video files using the previously trained object detector. This balancing is necessary as the best performance of neural networks can only be achieved if all training classes are of approximately the same size (Japkowicz and Stephen, 2002). Due to this requirement and the underrepresentation of LHD in the video data, it is possible to extract roughly 500 images per class and per 14-hour video on our dataset. The collection of all training images needs furthermore to be prepared a bit (AC 2a), so 5% - 10% of all images will be used as a validation set while the remaining 90% - 95% are the actual training set. Furthermore, to train the action classifiers for the binary classification task, the classes LHU and LHD need to be randomly merged, keeping in mind that LHU is the much more common posture and should therefore be overrepresented in the binary task in comparison to LHD. At this point, it is finally possible to train four EfficientNet-B3 CNNs, namely the single-frame classifier and the multiple-frame classifier for both (binary and total) classification tasks (AC 2b).

These first versions of the action classifiers are supposed to work quite decently on the videos used for the training, but it is likely that the classification accuracy is worse on different videos of the same animal in which the arrangement of the enclosure as well as the light conditions might be quite different. This turns out to be indeed a challenge as machine learning theory predicts that a deep learning system performs only well if the images in the training set are an almost uniform sample from the distribution of all possible images to be predicted and that, furthermore, such deep learning systems are brittle to distribution shifts (Quiñonero-Candela et al., 2008). For this reason, it seems sensible to reduce the latter. To this end, we adapt the classical OHEM to the setting at hand (AC 3) as follows. First, a fairly large number of momentarily not annotated video files will be predicted by BOVIDS (AC 3a). The accuracy of this prediction is supposed to be quite decent (at least 90%) as Hahn-Klimroth et al. (2021) already discussed. Therefore, BOVIDS provides an educated guess on labels of each time-interval of many video files that could not have been annotated manually. Based on those labels, one samples a decent number of images in almost balanced classes distributed over the whole observation time (AC 3b). These images are close to a uniform sample of the data on balanced classes and can therefore be referred to as “random” examples. These examples can now be evaluated by a human observer in a moderate amount of time (AC 3c). It is to be stressed at this point that a decent classifier is a critical ingredient: As the classes are highly unbalanced, random sampling without an educated guess would result in a set of images with almost no LHD, therefore, this simple process would not be possible to be used for generating a training set.

Besides mining such random examples, it is also possible to extract “hard” examples easily. In this contribution, a hard example is defined as an image for which either the certainty of classification by the single-frame action classifier is small or if it belongs to a time-interval of which the predictions of the single-frame and multiple-frame action classifier disagree. It is supposed that neural networks can be finetuned efficiently by hard examples (Felzenszwalb et al., 2010). Therefore, instead of only generating random samples distributed across the observation time, we can nudge the training set into a direction such that information from momentarily hard to classify data gets boosted.

Based on the human evaluation of the single images it is now possible to retrain the action classifiers on a much broader dataset that really represents the distribution to be classified. At this point, the training classes might get slightly unbalanced if the human annotation deviates strongly from the automatic one. In this case standard techniques like random upsampling might be considered (Branco et al., 2016) and are provided by BOVIDS. Once a decent object detector and a well-performing action classifier are generated, all data can be evaluated once more and the performance of BOVIDS can be measured.

2.3.5 BOV 4: Data prediction

The data prediction step consists of three major parts (DP 1 - DP 3) that are discussed in this section and read as

- DP 1: Prediction
  - P 1. Object detection phase
  - P 2. Action classification phase
  - P 3. Post-processing phase.
- DP 2: Data evaluation
- DP 3: Data presentation.

2.3.5.1 DP 1: The prediction pipeline

The system of Hahn-Klimroth et al. (2021) predicts a video file in three phases:

- P 1. Object detection phase
- P 2. Action classification phase
- P 3. Post-processing phase.

Those phases must not be confused with BOV 2 and BOV 3 that contain software to train the required deep neural networks while P 1 - P 3 are phases within the prediction pipeline of Hahn-Klimroth et al. (2021) that require the previously trained networks. In the following, those phases are briefly explained, and improvements and new features provided by BOVIDS in contrast to the original system are highlighted.

In the object detection phase (P 1), the system will first decompose a video file into so-called ‘time-intervals’. This is a discretization of the continuous data into packages of seven seconds each. More precisely, for each time-interval the prediction pipeline will collect four images. Then, the object detector is used to cut-out and identify the animal present in the images or, respectively, declare that no animal is present. While this step is governed by a Mask-RCNN network by Hahn-Klimroth et al. (2021) in the current version the architecture is changed to the much more recent yolov4 network which improves the classification accuracy (Bochkovskiy et al., 2020) and significantly speeds up the complete prediction pipeline by approximately 40% on the same hardware. The merit of this step is two-fold. First, as pointed out by Yosinski et al. (2014), it increases the similarity between images taken from different enclosures. This dramatically improves the chance of meaningful learning of the poses from various videos. Second, it is used to distinguish between distinct individuals within the same enclosure. This feature is a novelty of the present work, while it was previously reported as theoretically possible (Hahn-Klimroth et al., 2021). At the end of the object detection phase, each time-interval is represented in two ways for every individual: As a sequence of single images (single-frame) and additionally as one image in which these images are placed next to each other (multiple-frame encoded representation (Ji et al., 2013)).

The subsequent step, the action classification phase (P 2) to determine the behavioral classes, is a classical image classification task. For both, the single- and multiple-frame representations, this task is governed by two independently trained EfficientNetB3 CNNs per time-interval. The final prediction for any time-interval is calculated as the average over both outcomes. Hahn-Klimroth et al. (2021) already describe that the so-called "total classification" task (distinguishing Standing, LHU, LHD) might be much more difficult than the "binary classification" task (distinguishing Standing and Lying) and gives the possibility to map the final prediction from LHU and LHD to Lying. The approach of BOVIDS towards this binary task is slightly different. It is necessary to train a set of independent networks that purely govern this binary classification such that possible features can eventually be better learned.

To control classification flattering, Hahn-Klimroth et al. (2021) propose a set of post-processing rules (P 3) which are applied to the sequence of classifications of time-intervals. Those post-processing rules dismiss very short sequences of a specific action as those sequences are likely to stem from short periods of false classifications. In the current setting the set of post-processing rules is extended. It is now possible to handle flattering between Out and a specific behavior more smoothly to incorporate short periods in which the object-detector failed to detect or identify the present individual. Of course, such a post-processing step might dismiss short phases which are present in the data. Therefore, choosing an appropriate set of rules is a trade-off between a stronger methodological error (errors made by BOVIDS through misclassification of short events) and a systematic error (errors caused by dismissing short phases on real data). BOVIDS contains tools for a systematic study of both types of errors. Basically, it first applies the post-processing rules to the manually annotated data and analyses the accuracy as well as the number of dismissed short phases. If the systematic error is appropriate for the application, one can compare BOVIDS' prediction with the post-processed real data to describe the methodological error.

In the present work, the chosen set of post-processing rules varies significantly between the binary and the total classification task. Indeed, as the binary classification task is meant to study longer periods of Standing and Lying, phases up to 5 minutes are dismissed. Furthermore, in the total classification task, it is distinguished between adult Common Elands and non-adult Common Elands as the latter show shorter phases than the adult individuals. A detailed overview over the used post-processing rules can be found in Table 2 in the appendix.

2.3.5.2 DP 2: Data evaluation

As the prediction of a deep-learning based system works, in the end, as a black-box, it is very important to assure the quality of the prediction regarding all quantities of interest. Fortunately, a good testing set is already given by the manually annotated videos per individual. To quantify the accuracy of the prediction on the testing set, performance indicators from machine learning theory as well as biological key figures are evaluated by the following four quality criteria.

- QC 1. Analysis of the object detector per night ("detection density").
- QC 2. Accuracy and f-score as well as a comparison of the number of phases, the median phase length, and the overall percentage per activity class between BOVIDS' prediction and the manual annotation.
- QC 3. Number, length, and type of misclassified sequences.
- QC 4. Visual checking for outliers.

While QC 2 and QC 3 are quality criteria which can be only evaluated with respect to manually annotated videos, QC 1 and QC 4 can be applied to all predicted data.

In the first step (QC 1), the performance of the object detector should be checked in detail. It may happen that the object detector fails to detect the individual in certain videos quite often, which could be due to different light conditions or maybe because of heavy truncation. Of course, it is also possible that the individuals are Out for a longer period. To check the performance, BOVIDS outputs an overview that reports the percentage of detections of an individual by the object detector per video. If this value turns out to be noticeably low, one should check the original data to see if this low "detection density" can be explained.

Second, if the object detector works satisfactorily well and a good set of post-processing rules was set, the performance of the classification part of BOVIDS needs to be analyzed. To this end, it might be necessary to dismiss data with a significantly high amount of Out in the manually annotated nights. Once a satisfactory testing set is chosen, accuracy as well as f-score (QC 2) are standard tools to measure the performance of a deep learning system. The accuracy is defined as the percentage of correctly classified time-intervals by BOVIDS. While this is indeed an important key quantity, it does not describe BOVIDS' performance on underrepresented classes (like LHD) sufficiently. A more sensitive measure is the f-score, the harmonic mean between the positive predictive value (precision) and the sensitivity (recall) per class. Furthermore, it might be that the accuracy and the f-score are quite high but there is a lot of prediction flattering increasing the number of phases per activity class dramatically. Therefore, the latter should be compared between the post-processed manual annotation and BOVIDS' prediction. Further highly relevant biological quantities are the median phase length and the percentage per behavioral class. Thus, BOVIDS' prediction quality needs to be checked with respect to those quantities as well. Finally, it is important to understand which kind of misclassifications occur and to, potentially, derive patterns. To analyze QC 2 and QC 3, BOVIDS contains a tool that allows to report the accuracy, f-score, deviation in the number of phases as well as a detailed description of misclassified sequences.

If QC 1 - QC 3 are satisfactorily met, BOVIDS can be used to generate a final prediction of the unlabeled videos. Of course, QC 1 should be applied to unlabeled videos as well as it is a good indicator whether the object detector works well on a specific video. But even if the object detector detects an object quite frequently, it might happen that BOVIDS provides poor quality on a specific night if there are problems in the original data: for instance, individuals could be heavily truncated in a specific night. In those cases, it is reasonable to assume that the activity budget of the individual looks significantly different as in other videos which can be checked rather easily visually by searching for such outliers (QC 4). To this end, a short graphical representation of the activity budget in a video is generated by BOVIDS (see Figure 4) which can be used to identify those outliers. If the graphical representation of a night is conspicuous, one can check the original data on a sample basis to investigate BOVIDS' performance.

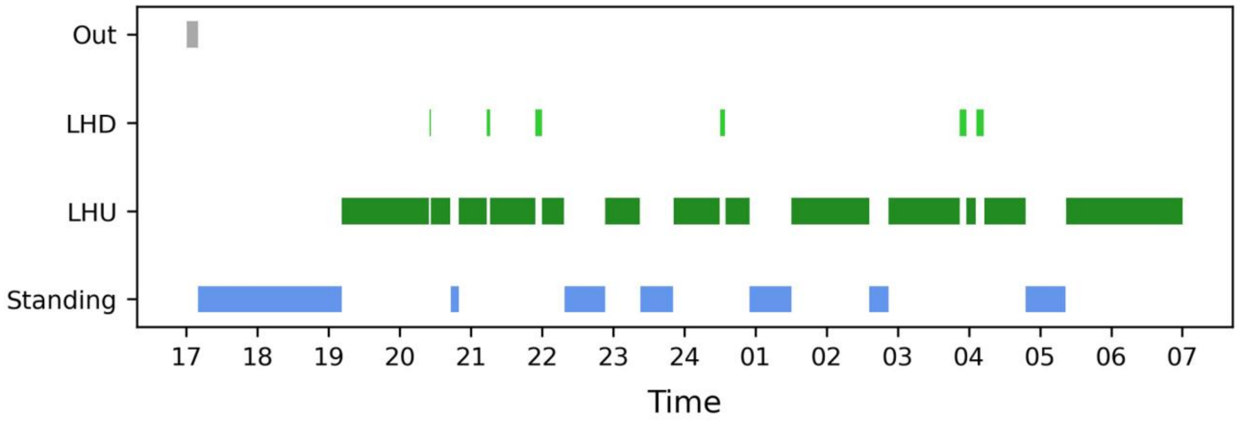

Figure 4. Example of one night of one Common Eland with the plotted phases of the three behavioral states of the total system given by BOVIDS to look for quality criteria QC 4.

2.3.5.3 DP 3: Data presentation

BOVIDS provides further functionalities to present the produced data elegantly which will be briefly described in this section and shown in more detail with the data of the case study in the results' section. Next to the graphical representation (see QC 4) of each night, BOVIDS produces a document that contains an overview about the most important statistical key quantities, for instance, the percentages of the single behaviors in the activity budget.

Finally, BOVIDS can be used to generate an overview about an individual's behavior over all evaluated nights or even about a species' average behavior over all nights of all individuals. The outcome is a table containing the important statistical key values, all the data of the single individuals and additional information to make data analysis with standard statistical software packages easy. Furthermore, first graphical representations of the nightly activity are given as can be seen in Figure 5.

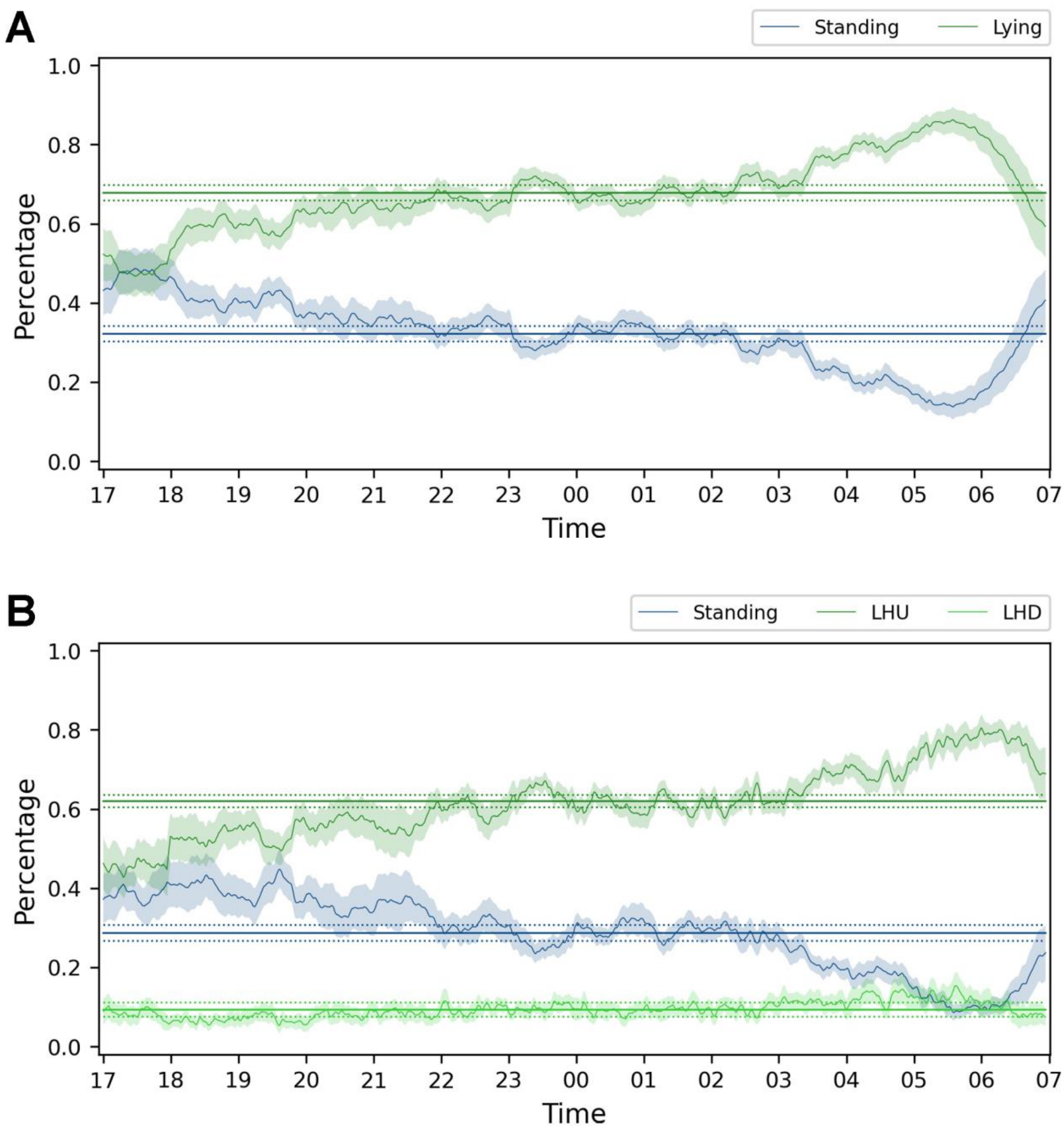

Figure 5. Timeline containing the percentage of all behavioral states and their means over all nights of all analyzed individuals of Common Elands. The visualization is smoothed by a rolling average over 3 minutes. (A) is the binary classification and contains 822 nights of 25 individuals, (B) is the total classification containing 589 nights of 16 individuals.

3 Results

3.1 BOVIDS' performance in the case study

This section is devoted to reporting the validity of post-processing rules and the quality criteria QC 1 - QC 4 in the case study. Subsequently, in the next section, the nocturnal behavior of the Common Elands is presented.

A set of post-processing rules can be considered as valid if the systematic error induced by these rules is negligible for the quantities of interest. On the dataset at hand and in both classification tasks, the accuracy of the post-processed data ranges from 99.6% to 100% and even the f-score of all activity classes lies constantly over 99.2%. Accordingly, the percentage per night per individual of all behavioral classes under both classification tasks are approximated up to an error of 0.02% in the worst case. Moreover, the average median phase length per individual is overshot by 21s of 1796s (Standing), 34s of 1375s (LHU) and 24s of 322s (LHD) in the total classification task while those values are 130s of 1834s (Standing) and 239s of 4226s (Lying) under binary classification. The number of phases per activity class is underestimated, more precisely, the mean deviation over all individuals is -0.29 of 8.2 (Standing), -1.02 of 23.0 (LHU) and -0.67 of 14.6 (LHD) in the total classification task while it is -1.4 of 8.9 (Standing) and -0.9 of 8.5 (Lying) in the binary classification system.

To analyze the quality criteria, the predictions of BOVIDS are compared to the manually annotated and post-processed nights. All nights in which individuals were at least 20% of the time Out, either by BOVIDS' prediction, or, if manually annotated by the humans' prediction, were dismissed as such nights do not yield good evidence on the individual's activity budget. Thus, the quality criteria are only analyzed for the remaining nights. The results of all quality criteria are presented in this section.

In the analysis of the accuracy (QC 2) of BOVIDS' prediction with respect to the manually annotated post-processed data, the following results are found. The median accuracy per night lies at 99.4% with a 0.25-quantile of 99.1% and a 0.75-quantile of 99.4% in the total classification task. Furthermore, the median f-scores turn out to be 99.6% (Standing), 99.5% (LHU) and 96.3% (LHD) with minima 94.4% (Standing), 95.4% (LHU) and, respectively, 93.2% (LHD). In the binary classification task, the corresponding values read as follows. The median accuracy is 99.8% with a 0.25-quantile of 99.4% and a 0.75-quantile of 99.8% while the f-scores are at least 93.1% (Standing) and 97.1% (Lying) with a median of 99.5% and, respectively, 99.8%. Furthermore, the percentage of each behavioral class per individual is approximated up to at most 0.03% in both classification tasks. In the total classification system, the mean deviation in the number of phases is 0.34 of 7.9 (Standing), 0.53 of 22.0 (LHU) and 0.37 of 13.9 (LHD). The values in the binary classification task are 0.05 of 7.5 (Standing) and 0.03 of 7.6 (Lying). Finally, the median phase length per individual is underestimated by -22.6s of 1817.6s (Standing), by -117.0s of 1409.9s (LHU) and -1.8s of 345.6s (LHD) in the total classification task. In the binary classification system, those values turn out to be -2.87s of 1970.9s (Standing) and -14.7s of 4464.5s (Lying).

The next quality criteria to analyze is the number, length, and type of misclassified sequences (QC3). In the total classification task, we find, overall, 179 misclassified sequences in 62 nights (thus, on average, 2.9 sequences per night). Out of 179 sequences, 49 sequences are misclassifications between a real behavior and being Out and in 65 cases, BOVIDS predicted LHD while the actual behavior was LHU. The remaining 65 sequences were mostly short confusions between Standing and LHU. In contrast, in the binary classification task, there are 181 misclassified sequences in 170 nights (on average 1.1 sequences per night) out of which 78 are confusions between a behavioral class and Out, in 78 cases, BOVIDS predicts Standing while the human label is Lying and in 27 cases vice versa. Furthermore, out of the 181 sequences, 46 misclassifications are sequences of length at most 1 minute and 47 additional misclassifications are below 5 minutes.

Quality criteria QC 1 and QC 4 are with respect to all predicted nights. Hereby, QC 1 checks the performance of the object detector. The detection density per individual ranges from 87.2% to 100% while its median turns out to be 99.8% with a 0.25-quantile of 97.5% and a 0.75-quantile of 100%. To analyze the last quality criteria (QC 4), namely, to look for apparent outliers, BOVIDS creates one plot per predicted night (for the binary and for the total classification task respectively) representing the timely course of the behavioral phases (see Figure 4). There are few apparent outliers on data which was not manually labeled, and the automatic annotation was checked randomly. In most cases, it was found that BOVIDS' prediction is correct even if it seemed to be suspicious. The observed misclassifications during this step fit exactly into the description of the errors in QC 3 and the frequency is comparable.

3.2 The nocturnal behavior of Common Elands

The data presentation tools of BOVIDS give a first visual result regarding the relative distribution of the behavioral states, their means over all nights, and the rhythm of phases of behavioral states (see Figure 5). The underlying data is normalized to the behavioral states excluding Out. The optically conjectured increase of Lying over the night between 19:00 and 06:00 in the binary classification task is confirmed by a linear regression ($R^2 = 0.799$ and $p < 0.0001$). In addition to the visual representation, BOVIDS' output consists of tables, including a summary table for every individual containing relevant statistical key values as well as a list of number of phases, durations, and the percentage of behaviors per night. This significantly facilitates the creation of an activity budget (see Figure 6) and provides a first insight into the nocturnal behavior of Common Elands. The graphical representation yields to the conjecture that there might be differences in the total duration of the behaviors per night between certain groups of individuals which are tested rigorously in the following.

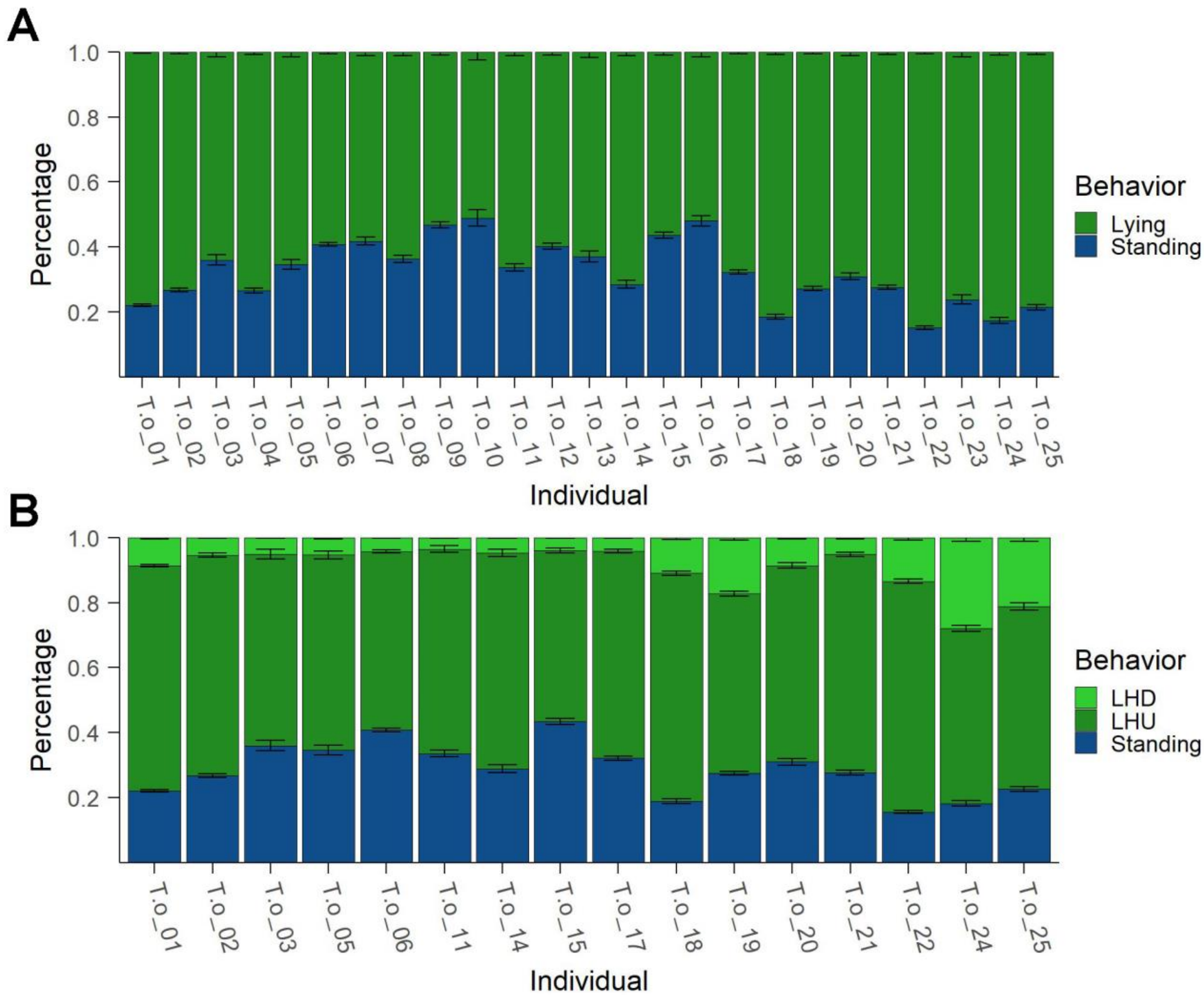

Figure 6. Activity budgets of all analyzed Common Elands: (A) in the binary classification with 822 nights of 25 individuals, (B) in the total classification with 589 nights of 16 individuals. *T.oryx_01* to *T.oryx_05* are male adult individuals and *T.oryx_06* to *T.oryx_17* female adult individuals while *T.oryx_18* to *T.oryx_21* are subadult and *T.oryx_22* to *T.oryx_25* are young individuals.

The data with respect to Standing and LHU can be assumed to be normally distributed (p_Standing = 0.9524 and p_LHU = 0.2715) while the total duration per night of LHD deviates significantly from normality (p_LHD = 0.0015) and is transformed. First, adult male and adult female individuals are compared to investigate sex differences. Afterwards, age specific analyses' will be conducted within the group of female individuals as there is only one non-adult male individual in the sample. To investigate the differences based on sex and to account for possible influences by the housing conditions, a two-factor ANOVA is conducted with the factors keeping zoo and sex between the adult animals for each behavior of the total classification system ($n = 9$ individuals with 328 nights consisting of 4 males with 151 nights and 5 females with 177 nights). The holding zoo can be withdrawn as a significant factor ($p > 0.37$), but the sex has a significant influence on LHD ($p = 0.0281$) whereby the males' values exceed the females', see Figure 7 (A). Finally, a two-factor ANOVA with factors keeping zoo and age within all female individuals in the total classification system ($n = 11$ individuals with 411 nights consisting of 3 young with 118, 3 subadults with 116 and 5 adults with 177 nights) is conducted. Again, the holding zoo can be withdrawn as a factor ($p > 0.58$), but the age influences the total duration of Standing (p_young-adult = 0.0038) and LHD (p_young-adult = 0.0009; p_subadult-adult = 0.0136) significantly as a corresponding post-hoc analysis verifies. Hereby, non-adult individuals spend more time on LHD than adult ones, whereby adult ones spend more time Standing, see Figure 7 (B). While the age comparison could only be carried out for female individuals, it is an advantageous circumstance that one individual could be recorded once as the subadult male individual (*T.oryx_18*) and moved during the observation phase to a different zoo in which it was observed as an adult male (*T.oryx_01*). This allows for a direct comparison of the behavior between the subadult and adult age of this individual as the husbandry conditions in the zoos studied were already considered negligible. An unpaired t-test shows that the total amount of Standing ($p < 0.0001$) and LHD ($p = 0.0001$) differs significantly between the two observation periods of this individual, confirming the previously found results in differences due to age.

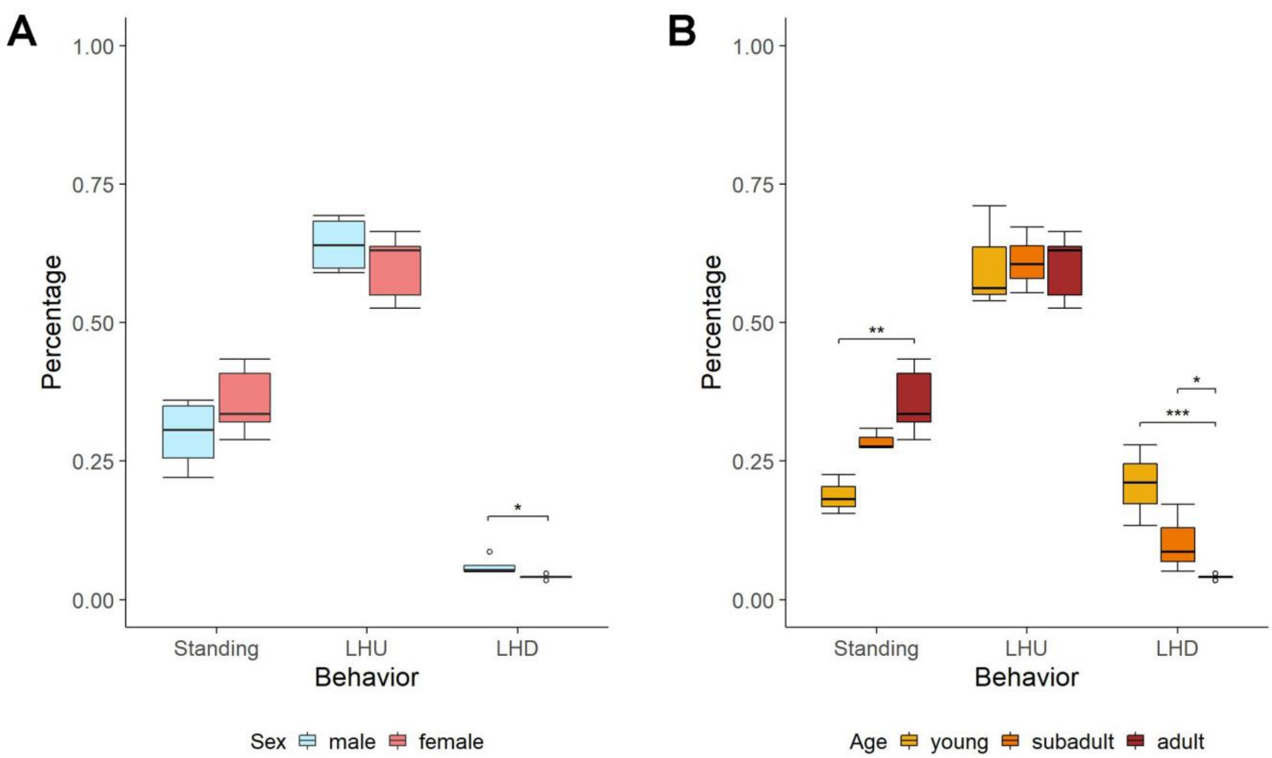

Figure 7. Comparison with respect to the total duration of each behavior per night in the total system. (A) Sex comparison (with n = 9 individuals with 328 nights, consisting of 4 males with 151 nights and 5 females with 177 nights) in which significant differences in LHD ($p = 0.0281$) arise. (B) Age comparison with (n = 11 individuals with 411 nights, consisting of 3 young with 118, 3 subadults with 116 and 5 adults with 177 nights) that yields to significant differences in Standing (p_young-adult = 0.0038) and LHD (p_young-adult = 0.0009; p_subadult-adult = 0.0136).

A second variable of interest is the length of each behavioral phase. Regarding this quantity, the binary classification system (Standing and Lying) was used for the analysis as well as the duration of LHD from the total classification system as one Lying phase might be interrupted by several events of LHD. A Wilcoxon test reveals that there are significant differences ($p = 0.0003$) in the median length of phases per individual within Lying between males and females (n = 17 individuals with 539, consisting of 5 males with 179 nights and 12 females with 360 nights). For this reason, these two groups were analyzed separately. Within the females (n = 19 individuals with 613 nights, consisting of 4 young with 137 nights, 3 subadults with 116 and 12 adults with 360 nights), a post-hoc analysis shows significant differences in the median duration of the Standing phases between young and adult individuals (p_Standing = 0.0033) and no significant differences between young and subadult animals (p_Standing = 0.1143, p_Lying = 0.629). Therefore, a detailed analysis is made after splitting into three categories adult male, adult female, non-adult (young and subadult) individuals. Figure 8 visualizes the distribution of the phase length regarding these categories. In median, the adult males exhibit the longest Lying phases with 89.6 minutes, followed by the non-adult animals (78.5 minutes) while the females show, with 59.3 minutes, the shortest Lying phases. While this is also true for the first and third quartile, the longest Lying event is achieved by the non-adults with 369.7 minutes. Within Standing, non-adult individuals seem to show a shorter median phase length (21.2 minutes) than adults (35.5 female, 30.8 male). With respect to phases of LHD, adult male individuals and non-adult individuals show, with a median value of 4.6 minutes and, respectively, 4.4 minutes a slightly longer duration than adult females with a median of 3.7 minutes. Nevertheless, the longest observed phase of LHD was by non-adult individuals (47.8 minutes) followed by the male adults (32.9 minutes) and the female adults (14.8 minutes).

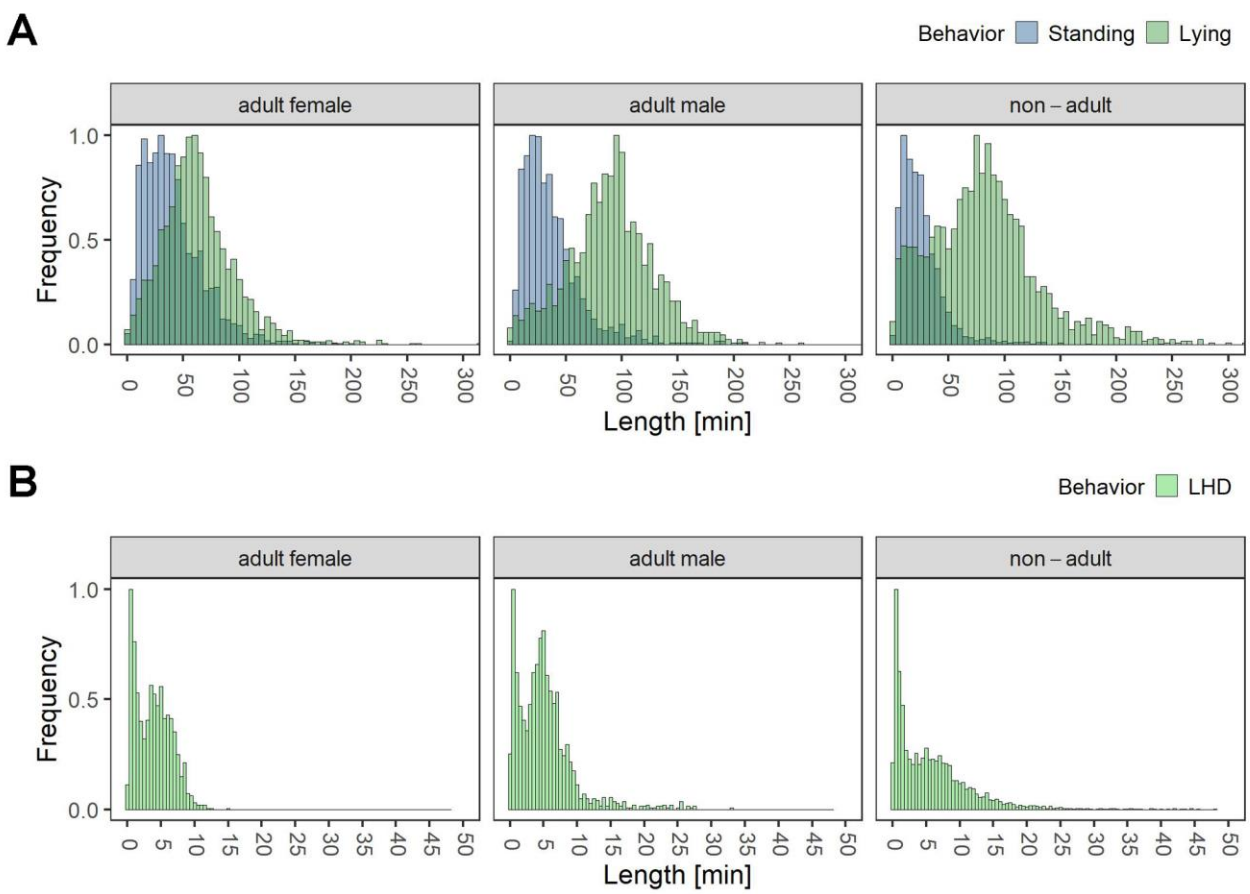

Figure 8. (A) For all 25 Common Elands is the distribution of the length of phases in minutes of Standing and Lying from the binary classification task plotted and the animals are classified as adult male (n = 5 individuals with 179 nights), adult female (n = 12 individuals with 360 nights) and non-adult animals (n = 8 with 280 nights). (B) Only the 16 Common Elands evaluated by the total classification system are used. The length of phases in minutes of LHD are plotted and the animals are classified as adult male (n = 4 individuals with 151 nights), adult female (n = 5 individuals with 177 nights) and non-adult animals (n = 7 individuals with 261 nights).

Beside the length of the phases, the number of phases per night is also an interesting parameter. Figure 9 visualizes the number of Lying phases (binary classification system) as well as the number of LHD phases (total classification system). Note that the number of Standing phases equals the number of Lying phases ± 1. The above illustration highlights the different age categories of young, subadults and adults, with sex being distinguished in the adult category. The phases in Lying (see Figure 9 (A)) appear to be constant across individuals and differences between sex or age groups are not evident. The situation is different when it comes to LHD, where the young animals have a significantly higher number of phases than the adults. The subadults tend to have slightly more LHD phases than the adults, but they are already closer to the values of the adults than to those of the young.

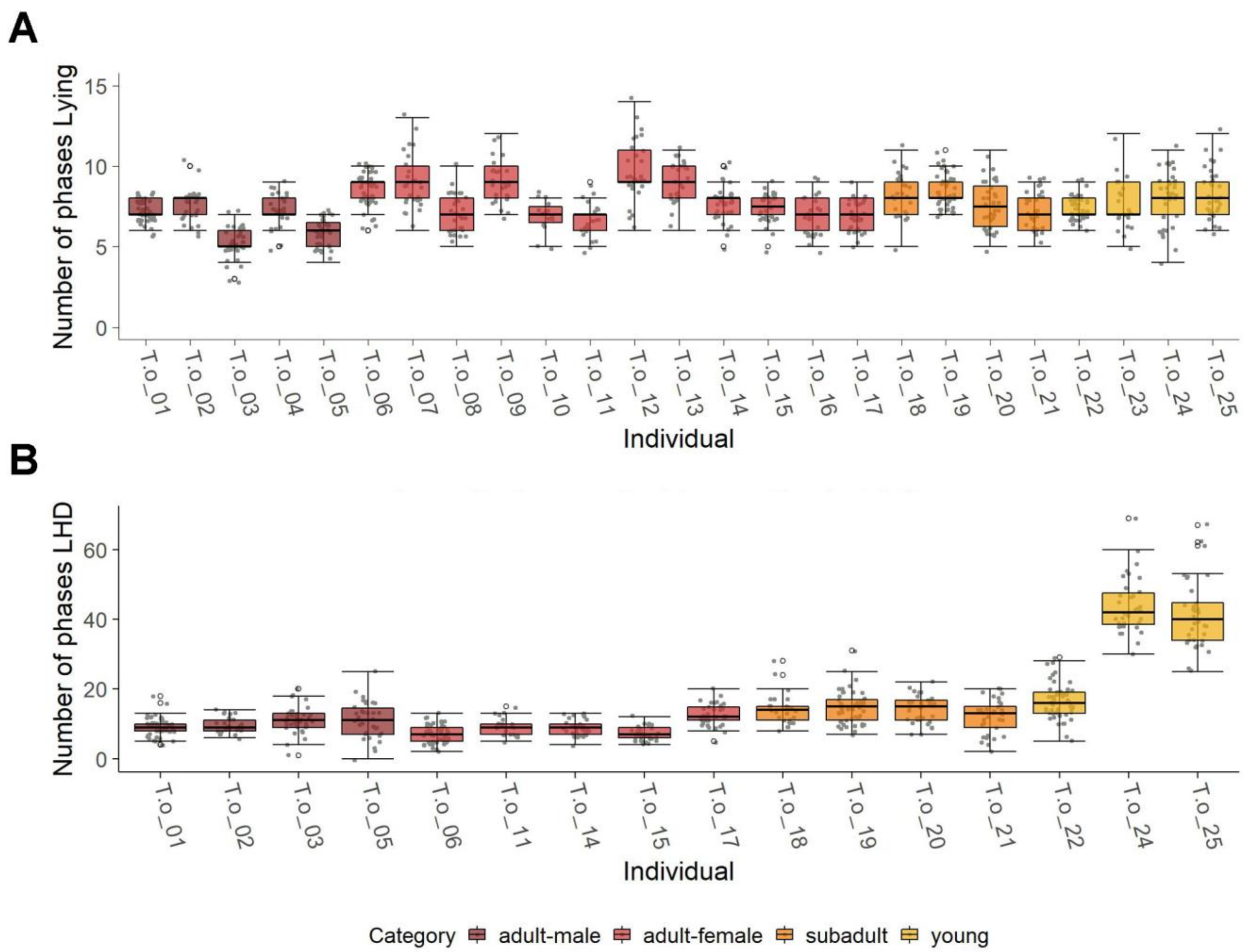

Figure 9. Number of phases for every individual marked are the groups adult male, adult female, subadult and young for (A) Lying and (B) LHD.

4 Discussion

4.1 BOVIDS

4.1.1 Performance in the case study

In this section, the validity of the post-processing rules as well as the four quality criteria are discussed. As can be seen in section *BOVIDS' performance in the case study*, only very few activity phases are dismissed on manually annotated nights when the selected post-processing rules are applied. Furthermore, both the accuracy and the f-scores are close to 100%, so that overall, the set of post-processing rules seems to be valid from a computer science point of view. Further, the percentage of each behavioral class is very well approximated in both classification tasks, so that no mentionable errors occur. Not very surprisingly, the post-processed data contains few phases less and slightly longer median phase lengths as very short events are dismissed, so the post-processing rules imply almost no bias in the real data. These values are of course a bit higher in the binary classification task, since longer phases up to five minutes are not considered. But firstly, even this choice does not imply much bias in the data, and secondly, the few short events of Standing and Lying do not significantly affect the animals' rhythms. Of course, neglecting the short events also increases the median phase length. However, this happens only very moderately, by a factor of between 5.6% (Standing) and 7.5% (LHD). It will be seen later that the methodological error will underestimate those quantities with respect to the post-processed data slightly. Therefore, the errors partly account for each other.

The object detector seems to work very well (QC 1) as the median object detection density is very high. On nights with a lower detection density, the video material was checked manually, and it can be observed that the individuals were mostly Out if the object detector did not find them, or only small parts are visible at the border of the video recording.

Subsequently, quality criteria QC 2 and QC 3 are discussed. Since the number of phases per activity class and the phase length analysis refer to Standing and Lying from the binary classification task as well as LHD from the total classification task, the discussion focuses on the reliability of these quantities. Overall, the accuracy and the f-score of BOVIDS' prediction are very high for machine learning based predictions. Recent studies on comparable hard data, such as that of Porto et al. (2013) on the discrimination of Standing and Lying behavior on video recordings of cows in stables, achieve an average accuracy of 92%. Our accuracies of 99.8% in the binary classification task and 99.4% in the total classification task clearly exceed this value. Furthermore, even the median f-score of the highly underrepresented class LHD is, with 96.4%, astonishingly high for a deep learning system. These values directly show that the percentage of each behavioral class is predicted very accurately and that there is no methodological bias in the expected activity budget.

Moreover, video action classifiers tend, normally, to so-called classification flickering, thus very short phases of misclassifications which do not really influence the accuracy and the f-score of the prediction system but have huge influence on the number of phases per activity. The post-processing rules are meant to take care of this effect (Hahn-Klimroth et al., 2021). The results show that BOVIDS succeeds in underestimating or overestimating the number of phases per activity class only very slightly on average. More precisely, the number of LHD phases is overestimated by 2.7% on average and the number of Standing and Lying phases is only overestimated by less than 1%. The median phase length is approximated very accurately as well, as it is only underestimated by at most 0.5% on average. It can be noted that even in enclosures containing two different individuals, BOVIDS' prediction does not become significantly worse. This has two reasons: First, and most importantly, the used object detector seems to be able to discriminate between two individuals very accurately. Secondly, the action classifier seems to be very robust against truncation effects when, for example, the bounding boxes of the two animals overlap.

In summary, the activity budget per night is predicted without any bias, as expected, while the median phase length per activity class is overestimated due to post-processing rules by a moderate factor of no more than 7.0%. Thus, the automatic prediction is very precise with respect to the post-processed data. Furthermore, the overall accurate description of the three poses Standing, LHU and LHD by BOVIDS can be seen in connection with the types of misclassifications occurring on the testing data. All misclassifications between Out and a real activity class are due to heavy truncation or occluding effects in which a human annotator might see hooves or small parts of the animal and is able to safely infer the behavior, but a machine cannot. In this case, it is favorable if the object detector does not find the animal in the first place. Furthermore, almost all misclassifications between LHU and LHD can be explained by the fact that Common Elands show, from time to time, a grooming behavior at their hind leg which is, on a single image, hard to distinguish from LHD. Such errors need, of course, to be considered and analyzed, but do not seem to be fixable by more training data or fine-tuning the networks if the input data format does not change significantly. As mentioned earlier, the median phase length as well as the median number of phases per night are only slightly overestimated. In the binary classification task, there are some short misclassifications with respect to the post-processed data less than five minutes in length. These errors are just delayed transitions between the behavioral states due to, for instance, the applied rolling average during post-processing. Therefore, these misclassifications neither influence the number of phases of Standing and Lying nor the animal's rhythms, but only slightly change the total duration of a specific phase. Finally, there are few misclassifications that are, probably, unavoidable in a deep learning classification task. Of course, accuracy can, in principle, always be improved by additional rounds of example mining and fine-tuning the action classifiers, but it is questionable whether an even higher median accuracy as 99.4% can be reached on a three-classes classification task.

A natural question, of course, is how well the findings from the test series can be generalized to unseen data of the same enclosures. Recall that the action classifiers are, in the end, trained on a random collection of images over the whole observation time due to offline hard example mining. Therefore, the testing set can be considered an almost random sample which includes a few more difficult images as expected on a random balanced sample. Thus, the analysis of the performance on the manually annotated nights (the testing set) yields a very good approximation of the overall performance. This claim is also supported by the analysis of QC 4. The type and frequency of errors on randomly selected, non-manually annotated nights were found to be comparable to those in the test set.

Finally, even if BOVIDS makes a small number of mistakes that would not occur if a trained observer manually annotated the data, the very large dataset overcompensates those few errors. Another approach to generating a large dataset is to have different, probably untrained, human observers annotate a comparable number of nights. Apart from the much higher cost, it is supposed that the inter-observer reliability might be worse than the reliability of BOVIDS. Overall, our findings show that BOVIDS performs very accurately in the case study and its predictions can be safely used to generate a large amount of annotated data, which would not have been easily possible without automation.

4.1.2 Universality, limitations, and extensions

A major strength of BOVIDS might be its adjustability to different settings. If the three positions Standing, LHU and LHD need to be detected from video files, the system can be used on data of ungulates. BOVIDS is tested extensively on the data of Common Elands and other African bovids stemming from various zoo enclosures. It is therefore reasonable to assume that, given sufficient training material, its performance is equally high under varying conditions. For instance, it is likely to perform well in the observation of various ungulates of different sizes from multiple continents in zoo enclosures or the analysis of livestock's behavior in stables. Since the present data are recorded in very different enclosures with partly high degrees of truncation and background noise, BOVIDS might perform well in outdoor enclosures if the camera installment is flawlessly possible in the sense that the whole outdoor enclosure can be recorded which would be a large step towards observations in the ungulates' natural habitat. A further research direction would be the analysis of BOVIDS' performance on data of larger groups of ungulates. While technically the detection of individuals works the same, it is clearly a much more difficult task to distinguish many individuals from each other than it is to identify two individuals reliably. Finally, on the technical side, it might be tempting to extend BOVIDS' functionality. For instance, if individual detection fails in large groups, one could implement a scan-sampling method that allows to at least report an average behavior of all the individuals.

4.2 The nocturnal behavior of Common Elands

A first finding is that independent from the factors age, sex, and keeping zoo, all individuals exhibit a higher percentage of Lying than Standing during the night. As the night progresses, the percentage of Lying increases significantly. This is in line to findings of similar studies on African Elephants (*Loxodonta africana*), Blue Wildebeest (*Connochaetes taurinus*) or Arabian Oryx (*Oryx leucoryx*), where the observed animals also show most of the sleeping behavior or inactivity in the second part of the night (Gravett et al., 2017; Davimes et al., 2018; Malungo et al., 2021).

When looking at LHD, it should be noted that this most likely corresponds to the typical REM (rapid eye movement) sleep posture. As mentioned in the ethogram section, a behavioral component to recognize REM sleep is the head being down due to postural atonia (Lima et al., 2005; Zepelin et al., 2005). In this study, we use this characteristically REM sleep posture to determine REM sleep. This approach is in line with the study by Zizkova et al. (2013) on Common Elands and the study by Ternman et al. (2014) on cows, which shows that REM sleep can be detected with sufficient certainty based on behavioral surveys. This procedure is also supported by a study on Lesser Mouse-deer (*Tragulus kanchil*), which shows that REM sleep can be divided into two categories, one of which is the most common, where the head lies down most of the time, making this a valid indicator to recognize REM sleep in behavioral studies (Lyamin et al., 2021).

Sex has been found to have an influence on the total amount of LHD during the night. Here, the adult females sleep slightly longer than the adult males, a fact which is also known across multiple phylogenetic states, for birds and mammals (Cajochen et al., 2006; Steinmeyer et al., 2010; Rattenborg et al., 2017). However, other studies show that there are no sex differences when individuals are similar-sized between the sexes, while dissimilar-sized animals should have differences (Ruckstuhl and Kokko, 2002). In Common Elands, males are larger than females (Leslie Jr, 2011; Myers et al., 2021), confirming the differences found between the sexes. In addition, Standing was found to increase with age. Interestingly, this finding is supported by the recording of a male individual at both subadult and adult age, which shows a significant increase in the total amount of Standing per night. Our results are in line with previous results on different mammals, as age is known to be an influencing factor for activity/rest cycles (Siegel, 2005; Ruckstuhl and Neuhaus, 2009; Steinmeyer et al., 2010). Moreover, age also influences REM sleep behavior in mammals and birds (Ruckstuhl and Kokko, 2002; Cajochen et al., 2006; Steinmeyer et al., 2010; Rattenborg et al., 2017). This effect was also observed in the Common Elands in this study, where the extent of LHD differs between the three age classes young, subadults and adults. A study on Giraffes (*Giraffa camelopardalis*) also shows that age and sex have an influence on the behavior Standing, while only age has an influence on REM sleep (Burger et al., 2021). The study by Burger et al. (2021) further reveals that housing conditions can be discarded as an influencing factor for both behaviors. These results correspond to the results in this study with Common Elands, where the keeping zoo and thus housing conditions can also be discarded as influencing factors. Of course, the factor housing condition consists of several factors as, among others, enclosure size, the presence or not of other types of animals in the vicinity or lighting conditions. While the recorded data does not allow to evaluate each possibly influencing factor individually, our study reveals that the sum of those effects is negligible and can be discarded.

Besides the total amount of time during the night, the duration of the single phases is also of interest. Here, the males differ from the females within Lying, whereby males show longer Lying phases than females. This fits with the result that adult males have a higher amount of LHD. Also, the age has an influence on the lengths of the phases. The non-adult animals show shorter periods of Standing and longer periods of Lying than the adult ones. This also matches with the results regarding the nocturnal activity budgets. Within LHD the number of phases vary between the different categories of individuals. The mean phase length of LHD in all adult Common Elands is 9.5 minutes on average, with females slightly below this at 8.8 minutes and males slightly above at 10.2 minutes. These phase lengths are consistent with those of male Arabian Oryx (*Oryx leucoryx*), which have a mean phase length of 7 ± 2 minutes in the dark in winter, and 10.5 ± 1.5 minutes over the 24 h cycle (Davimes et al., 2018).

Finally, also the number of phases is an interesting key figure in behavioral analysis. Within Lying and Standing it is thrilling that almost all animals show very similar numbers of phases. Here, of the 25 animals observed, 23 have a median between 7 and 9 phases per night with quite little variation per individual. The other two animals are moderate outliers downwards. Also, the mean lies between 6.6 and 9.1 within 22 individuals and within all individuals the SEM is at most 0.36 indicating a constant behavior within the single individuals. This suggests that certain rhythms are present and should be investigated in more detail in further analyses, because the course over the night also suggests certain rhythms. Within LHD, a different picture of the underlying distributions emerges. Here, the adult individuals show a lower proportion than the non-adult individuals, and within the non-adult individuals there are also strong differences between the young ones and the subadult ones. Only a few exceptions are to be recognized, which are explainable as follows. *T.oryx_22* is clearly different from the veined young and is closer to the values of the subadult individuals. However, *T.oryx_22* is also the oldest animal among the group of young ones. Furthermore, *T.oryx_17*, which is the oldest animal in the case study, has a higher median number of phases than the other adult animals, especially the female ones. Excluding these exceptions, young individuals have a median of 40-42 phases of LHD and subadults show 13-15 phases. In contrast, adult females have 7-9 phases of LHD and adult males 9-11 phases. This again indicates differences between the sexes and high similarities within each group of individuals. Again, it seems that certain rhythms are present depending on the sex and the age but being independent from the specific individual. This observation might be the starting point of a much more detailed analysis of rhythms in African ungulates' behavior.

5 Data Accessibility Statement

The python code is available at GitHub: *https://github.com/Klimroth/BOVIDS* and will, in case of publication, also be uploaded to Figshare.

6 Ethics Statement

The Common Eland's behavior was observed by videography such that the animals were disturbed as little as possible. This study was non-invasive as it was observational in nature and caused no undue harm to the Common Elands. All participating zoos supported this study.

7 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8 Author Contributions

JG conceptualized this project. PD acquired the funding. MH developed the deep learning software. JG collected, analyzed, and prepared the data. All authors contributed to the discussion and interpretation of the data, as well as the writing of the original draft.

9 Acknowledgments

We thank the Opel-Zoo Foundation Professorship in Zoo Biology of the *von Opel Hessische Zoostiftung* who funded the research leading to the results. The authors gained great support from directors and curators of the participating zoos (in alphabetical order): Allwetterzoo Münster, Erlebnis-Zoo Hannover, Opel-Zoo Kronberg, Zoo Dortmund, Zoom Erlebniswelt Gelsenkirchen. Also, the animal keepers of these zoos promoted this study mainly by assisting and allowing the data collection and providing information about the animals. Isabel Seyrling and Franziska Zölzer assisted in the installation of the cameras.

10 References

Bennie, J. J., Duffy, J. P., Inger, R., and Gaston, K. J. (2014). Biogeography of time partitioning in mammals. *Proc Natl Acad Sci U S A* 111, 13727–13732. doi: 10.1073/pnas.1216063110

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv [Preprint]. Available at: https://arxiv.org/pdf/2004.10934 (Accessed September 13, 2021)

Branco, P., Torgo, L., and Ribeiro, R. P. (2016). A Survey of Predictive Modeling on Imbalanced Domains. *ACM Comput. Surv.* 49, 1–50. doi: 10.1145/2907070

Brando, S., and Buchanan-Smith, H. M. (2018). The 24/7 approach to promoting optimal welfare for captive wild animals. *Behav Processes* 156, 83–95. doi: 10.1016/j.beproc.2017.09.010

Burger, A. L., Fennessy, J., Fennessy, S., and Dierkes, P. W. (2020). Nightly selection of resting sites and group behavior reveal antipredator strategies in giraffe. *Ecol Evol* 10, 2917–2927. doi: 10.1002/ece3.6106

Burger, A. L., Hartig, J., and Dierkes, P. W. (2021). Biological and environmental factors as sources of variation in nocturnal behavior of giraffe. *Zoo Biology* 40, 171–181. doi: 10.1002/zoo.21596

Cajochen, C., Münch, M., Knoblauch, V., Blatter, K., and Wirz-Justice, A. (2006). Age-related changes in the circadian and homeostatic regulation of human sleep. *Chronobiology International* 23, 461–474. doi: 10.1080/07420520500545813

Chakravarty, P., Cozzi, G., Dejnabadi, H., Léziart, P.-A., Manser, M., Ozgul, A., et al. (2020). Seek and learn: Automated identification of microevents in animal behaviour using envelopes of acceleration data and machine learning. *Methods Ecol Evol* 11, 1639–1651. doi: 10.1111/2041-210X.13491

Davimes, J. G., Alagaili, A. N., Bhagwandin, A., Bertelsen, M. F., Mohammed, O. B., Bennett, N. C., et al. (2018). Seasonal variations in sleep of free-ranging Arabian oryx (*Oryx leucoryx*) under natural hyperarid conditions. *Sleep* 41. doi: 10.1093/sleep/zsy038

Dell, A. I., Bender, J. A., Branson, K., Couzin, I. D., Polavieja, G. G. de, Noldus, L. P. J. J., et al. (2014). Automated image-based tracking and its application in ecology. *Trends Ecol Evol* 29, 417–428. doi: 10.1016/j.tree.2014.05.004

Eikelboom, J. A. J., Wind, J., van de Ven, E., Kenana, L. M., Schroder, B., Knegt, H. J. de, et al. (2019). Improving the precision and accuracy of animal population estimates with aerial image object detection. *Methods Ecol Evol* 10, 1875–1887. doi: 10.1111/2041-210X.13277

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 1627–1645. doi: 10.1109/TPAMI.2009.167

Friard, O., and Gamba, M. (2016). BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. *Methods Ecol Evol* 7, 1325–1330. doi: 10.1111/2041-210X.12584

Gerovichev, A., Sadeh, A., Winter, V., Bar-Massada, A., Keasar, T., and Keasar, C. (2021). High Throughput Data Acquisition and Deep Learning for Insect Ecoinformatics. *Front. Ecol. Evol.* 9, 309. doi: 10.3389/fevo.2021.600931

Gravett, N., Bhagwandin, A., Sutcliffe, R., Landen, K., Chase, M. J., Lyamin, O. I., et al. (2017). Inactivity/sleep in two wild free-roaming African elephant matriarchs - Does large body size make elephants the shortest mammalian sleepers? *PLoS One* 12, e0171903. doi: 10.1371/journal.pone.0171903

Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., et al. (2019). DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife* 8, 1–42. doi: 10.7554/eLife.47994

Groves, C. P., and Leslie Jr, D. M. (2011). “Family Bovidae (Hollow-horned Ruminants),” in *Handbook of the Mammals of the World: Hoofed Mammals*, eds. D. E. Wilson, and R. A. Mittermeier (Barcelona: Lynx Edicions), 444–780.

Hahn-Klimroth, M., Kapetanopoulos, T., Gübert, J., and Dierkes, P. W. (2021). Deep learning-based pose estimation for African ungulates in zoos. *Ecol Evol* 11, 6015–6032. doi: 10.1002/ece3.7367

Japkowicz, N., and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis* 6, 429–449. doi: 10.5555/1293951.1293954

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 221–231. doi: 10.1109/TPAMI.2012.59

Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S., and Branson, K. (2013). JAABA: interactive machine learning for automatic annotation of animal behavior. *Nat Methods* 10, 64–67. doi: 10.1038/nmeth.2281

Leslie Jr, D. M. (2011). “Family Bovidae (Hollow-horned Ruminants): Species accounts of *Taurotragus oryx*”, in *Handbook of the Mammals of the World: Hoofed Mammals*, eds. D. E. Wilson, and R. A. Mittermeier (Barcelona: Lynx Edicions), 617.

Lima, S. L., Rattenborg, N. C., Lesku, J. A., and Amlaner, C. J. (2005). Sleeping under the risk of predation. *Animal Behaviour* 70, 723–736. doi: 10.1016/j.anbehav.2005.01.008

Lürig, M. D., Donoughe, S., Svensson, E. I., Porto, A., and Tsuboi, M. (2021). Computer Vision, Machine Learning, and the Promise of Phenomics in Ecology and Evolutionary Biology. *Front. Ecol. Evol.* 9, 148. doi: 10.3389/fevo.2021.642774

Lyamin, O. I., Siegel, J. M., Nazarenko, E. A., and Rozhnov, V. V. (2021). Sleep in the lesser mouse-deer (*Tragulus kanchil*). *Sleep*. doi: 10.1093/sleep/zsab199

Malungo, I. B., Gravett, N., Bhagwandin, A., Davimes, J. G., and Manger, P. R. (2021). Sleep in two free-roaming blue wildebeest (*Connochaetes taurinus*), with observations on the agreement of polysomnographic and actigraphic techniques. *IBRO Neuroscience Reports* 10, 142–152. doi: 10.1016/j.ibneur.2021.02.005

Martin, P., and Bateson, P. P. G. (2015). *Measuring behaviour: An introductory guide*. Cambridge: Cambridge University Press.

Merrow, M., Spoelstra, K., and Roenneberg, T. (2005). The circadian cycle: daily rhythms from behaviour to genes. *EMBO reports* 6, 930–935. doi: 10.1038/sj.embor.7400541

Myers, P., Espinosa, R., Parr, C. S., Jones, T., Hammond, G. S., and Dewey., T. A. (2021). *The Animal Diversity Web*. https://animaldiversity.org (Accessed April 09, 2021)

Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., and Clune, J. (2021). A deep active learning system for species identification and counting in camera trap images. *Methods Ecol Evol* 12, 150–161. doi: 10.1111/2041-210X.13504

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., et al. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc Natl Acad Sci U S A* 115, E5716-E5725. doi: 10.1073/pnas.1719367115

Pereira, T. D., Tabris, N., Li, J., Ravindranath, S., Papadoyannis, E. S., Wang, Z. Y., et al. (2020). SLEAP: Multi-animal pose tracking. biorxiv [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/2020.08.31.276246v1 (Accessed September 13, 2021)

Peterson, R. A., and Cavanaugh, J. E. (2020). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics* 47, 2312–2327. doi: 10.1080/02664763.2019.1630372

Porto, S. M., Arcidiacono, C., Anguzza, U., and Cascone, G. (2013). A computer vision-based system for the automatic detection of lying behaviour of dairy cows in free-stall barns. *Biosystems Engineering* 115, 184–194. doi: 10.1016/j.biosystemseng.2013.03.002

Puschmann, W., Zscheile, D., and Zscheile, K. (2009). *Säugetiere: Zootierhaltung*. Tiere in menschlicher Obhut. Frankfurt am Main: Harri Deutsch.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., eds (2008). *Dataset shift in machine learning*. Cambridge, Mass. MIT Press.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical.

Rattenborg, N. C., La Iglesia, H. O. de, Kempenaers, B., Lesku, J. A., Meerlo, P., and Scriba, M. F. (2017). Sleep research goes wild: new methods and approaches to investigate the ecology, evolution and functions of sleep. *Philos Trans R Soc Lond B Biol Sci* 372. doi: 10.1098/rstb.2016.0251

Ruckstuhl, K., and Kokko, H. (2002). Modelling sexual segregation in ungulates: effects of group size, activity budgets and synchrony. *Animal Behaviour* 64, 909–914. doi: 10.1006/anbe.2002.2015

Ruckstuhl, K. E., and Neuhaus, P. (2009). Activity budgets and sociality in a monomorphic ungulate: the African oryx (*Oryx gazella*). *Can. J. Zool.* 87, 165–174. doi: 10.1139/Z08-148

Ryder, O. A., and Feistner, A. T. C. (1995). Research in zoos: A growth area in conservation. *Biodiversity and Conservation* 4, 671–677. doi: 10.1007/BF00222522

Sicks, F. (2016). REM sleep as indicator for stress in giraffes (*Giraffa camelopardalis*). *Mammalian Biology* 81, 16. doi: 10.1016/j.mambio.2016.07.052

Siegel, J. M. (2005). Clues to the functions of mammalian sleep. *Nature* 437, 1264–1271. doi: 10.1038/nature04285

Steinmeyer, C., Schielzeth, H., Mueller, J. C., and Kempenaers, B. (2010). Variation in sleep behaviour in free-living blue tits, *Cyanistes caeruleus*: effects of sex, age and environment. *Animal Behaviour* 80, 853–864. doi: 10.1016/j.anbehav.2010.08.005

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., and Packer, C. (2015). Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Sci Data* 2, 150026. doi: 10.1038/sdata.2015.26

Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., et al. (2013). Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res* 41, D1027-33. doi: 10.1093/nar/gks1155

Taipingeric (2020). *yolo-v4-tf.keras*. https://github.com/taipingeric/yolo-v4-tf.keras (Accessed September 13, 2021)

Tan, M., and Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ICLR*, 6105–6114.

Ternman, E., Pastell, M., Agenäs, S., Strasser, C., Winckler, C., Nielsen, P. P., et al. (2014). Agreement between different sleep states and behaviour indicators in dairy cows. *Applied Animal Behaviour Science* 160, 12–18. doi: 10.1016/j.applanim.2014.08.014

Tzutalin (2015). *LabelImg*. https://github.com/tzutalin/labelImg (Accessed September 13, 2021)

Valletta, J. J., Torney, C., Kings, M., Thornton, A., and Madden, J. (2017). Applications of machine learning in animal behaviour studies. *Animal Behaviour* 124, 203–220. doi: 10.1016/j.anbehav.2016.12.005

Walsh, B., Binding, S., and Holmes, L. (2019). While you were sleeping… *Zooquaria* 105, 28–29.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wu, Y., Wang, H., Wang, H., and Feng, J. (2018). Arms race of temporal partitioning between carnivorous and herbivorous mammals. *Sci Rep* 8, 1713. doi: 10.1038/s41598-018-20098-6

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Proceedings of the 27th International Conference on Neural Information Processing Systems* 2, 3320–3328.

Zepelin, H., Siegel, J. M., and Tobler, I. (2005). "Chapter 8 - Mammalian Sleep," in *Principles and Practice of Sleep Medicine*, eds. M. H. Kryger, T. Roth, and W. C. Dement (New York: Elsevier), 91–100.

Zizkova, K., Kotrba, R., and Kocisova, A. (2013). Effect of Changes in Behaviour on the Heart Rate and its Diurnal Variation in a Male and a Female Eland (*Taurotragus oryx*). *Agricultura tropica et subtropica* 46, 29–33. doi: 10.2478/ats-2013-0005

11 Appendix

11.1 Overview data

A detailed overview about the used data is given in Table 1**.** Hereby, for every individual the categories age, sex and the keeping zoo as well as the stabeling conditions are contained. The exact age of the observed individuals ranges from one month to 16.5 years categorized as follows: 'young' ranges from birth until the time of weaning with about 6 months, then the individuals become 'subadult' until sexual maturity with about 2 years of age and after that they are listed as 'adult'.

Table 1. The Common Elands observed in this study and their individual factors age (categorical: young, subadult and adult) and sex. Further, the housing zoo and the given stabeling conditions (standing single or together), are contained. The duration gives the recording start and end time and the totally recorded number of nights as well as the manually annotated number of nights are listed, if nights had to be removed because of an object detection density score smaller than 80% the used number of nights are listed with the real number of nights in parentheses. Finally, the number of pictures describes the number of annotated images in the object detection training set after OHEM. Observe that *T.oryx_01* and *T.oryx_18* is the same individual recorded at different times after moving from one zoo to another. Also, it is marked if the individuals are evaluated with the total or binary classification system.

Individual	Age	Sex	Keeping	Stabeling	Nights	Duration	Nights per hand	Pictures	Binary	Total
T.oryx_01	adult	m	Zoo_1	single	49	17-7 h	2	404	x	x
T.oryx_02	adult	m	Zoo_4	single	29	17-7 h	10	544	x	x
T.oryx_03	adult	m	Zoo_3	single	38	18-7 h	2	517	x	x
T.oryx_04	adult	m	Zoo_5	single	28	17-7 h	15	860	x	--
T.oryx_05	adult	m	Zoo_2	single	35	17-7 h	4	519	x	x
T.oryx_06	adult	f	Zoo_1	single	49	17-7 h	2	404	x	x
T.oryx_07	adult	f	Zoo_4	single	29	17-7 h	10	487	x	--
T.oryx_08	adult	f	Zoo_4	single	29	17-7 h	10	519	x	--
T.oryx_09	adult	f	Zoo_4	single	29	17-7 h	10	504	x	--
T.oryx_10	adult	f	Zoo_4	single	15	17-7 h	10	512	x	--
T.oryx_11	adult	f	Zoo_3	single	21	18-7 h	2	550	x	x
T.oryx_12	adult	f	Zoo_5	single	28	17-7 h	11	513	x	--
T.oryx_13	adult	f	Zoo_5	single	28	17-7 h	14	541	x	--
T.oryx_14	adult	f	Zoo_2	together	35	17-7 h	2	604	x	x
T.oryx_15	adult	f	Zoo_2	together	34	17-7 h	2	604	x	x
T.oryx_16	adult	f	Zoo_4	single	25	17-7 h	10	557	x	--
T.oryx_17	adult	f	Zoo_3	single	38	18-7 h	2	511	x	x
T.oryx_18	subadult	m	Zoo_5	together	27 (28)	17-7 h	17 (18)	502	x	x
T.oryx_19	subadult	f	Zoo_1	together	49	17-7 h	2	636	x	x
T.oryx_20	subadult	f	Zoo_2	single	34	17-7 h	4	519	x	x
T.oryx_21	subadult	f	Zoo_2	single	33	17-7 h	4	519	x	x
T.oryx_22	young	f	Zoo_1	together	49	17-7 h	2	636	x	x
T.oryx_23	young	f	Zoo_5	together	22 (28)	17-7 h	15 (18)	502	x	--
T.oryx_24	young	f	Zoo_2	together	35	17-7 h	2	604	x	x
T.oryx_25	young	f	Zoo_2	together	34	17-7 h	2	604	x	x

11.2 Post-processing rules

This section contains the post-processing rules applied to BOVIDS' prediction for both classification tasks. With respect to the total classification task, different sets of rules are applied for adult Common Elands and non-adult Common Elands, because non-adult individuals show shorter phases.

The order of the applied rolling average varies between the three sets of rules. A higher order reduces flickering but is likely to dismiss (very) short events. Therefore, the order of the rolling average was set to 3 in the total classification task for non-adult individuals, to 4 in the total classification task for adult individuals and to 5 in the binary classification task.

Regarding dismissing short phases, the quantity "minimum length" is introduced followed by a three-character code. If this code is XYZ, this is meant to be read as follows. Suppose a phase of behavior Y lies in between a phase of behavior X and behavior Z, then the event will be dismissed (marked as X) if it consists of less time-intervals than indicated by the minimum length of XYZ. In those codes, Standing is abbreviated to "A", LHU to "L" and LHD to "S" in the total classification task. In the binary classification task, "A" means Standing and "L" means Lying. "O" stands for Out in both tasks. *X* is meant to be read as any combination YXZ where Y and Z do not equal X. The applied rules of dismissing short phases can be found in Table 2.

Regarding the special state Out, the post-processing rules are a bit more elaborated. If flickering between Out and a real behavioral state occurs, this is very likely due to a failure of the object detector if an animal is occluded or truncated. Therefore, if a sequence of a specific behavioral state X (Standing, Lying, LHU or LHD) is interrupted by phases of Out, the Out phases are dismissed under the following conditions. First, each single phase of Out must be shorter than 27 time-intervals (total) or 135 time-intervals (binary). Second, the total percentage of X in the sequence needs to exceed 20%.

Table 2. Overview about the minimum length a specific behavioral phase needs to have in order not to be dismissed in the post-processing step. The value is to be read as time-intervals where 1 time-interval consists of 7 seconds. Standing is abbreviated to "A", LHU to "L" and LHD to "S" in the total classification task. In the binary classification task, "A" means Standing and "L" means Lying. "O" stands for Out in both tasks.

Behavior Code	total adult	total non-adult	binary
SLS	3	2	---
SLA	3	3	---
ALS	3	3	---
ALA	6	6	45
OLA	6	6	45
OLS	6	6	---
ALO	6	6	45
SLO	6	6	---
SAS	25	6	---
SAL	25	6	---
LAS	25	6	---
LAL	25	6	45
LAO	25	6	45
OAL	25	6	45
OAS	25	6	---
SAO	25	6	---
ASA	9	9	---
ASL	6	6	---
LSA	6	6	---
LSL	2	2	---
LSO	9	9	---
OSL	9	9	---
ASO	9	9	---
OSA	9	9	---
O	9	9	45