Male aggressive interactions and courtship behaviour identified in a wild vulturine guineafowl population from accelerometer data using deep learning

ESS 511 Master's Thesis

Author
Tobias Kuster
16-700-791

Supervised by
Prof. Dr. Damien Farine (damien.farine@ieu.uzh.ch)
Dr. André Marques Condeço Ferreira (andre.marquescondecoferreira@uzh.ch)
Dr. Charlotte Christensen (charlotte.christensen@uzh.ch)

Faculty representative
Prof. Dr. Robert Weibel

22.05.2024
Department of Geography, University of Zurich

Male aggressive interactions and courtship behaviour identified in a wild vulturine guineafowl population from accelerometer data using deep learning

Master Thesis
Tobias Kuster

Supervisors:
Prof. Dr. Damien Farine
Department of Evolutionary Biology and Environmental Studies, UZH

Prof. Dr. Robert Weibel
Department of Geography, UZH

Co-Supervisors:
Dr. Charlotte Christensen

Dr. André Marques Condeço Ferreira
Department of Evolutionary Biology and Environmental Studies, UZH

Summary

In the realm of studying the intricacies of the natural world, few fields captivate the curiosity, fascination, imagination, and intellect quite like the study of animal behaviour. From the earliest observations of ancient philosophers to today's cutting-edge research, the exploration of how animals think, feel, make decisions, and interact with their environment has remained a cornerstone of scientific inquiry. Understanding animal behaviour offers insights into the adaptive strategies of organisms for survival, competition, and reproduction, to thrive in a dynamic, ever- but increasingly fast-changing world. Some social interactions are difficult to observe and study as they can be rare and short, but very important events in group-living animals. They contribute significantly to the social environment and are in interdependence with ecological dynamics and evolutionary adaptations which have fitness consequences. From the intricate dance of courtship rituals, which secure mating, to the complex dynamics of social hierarchies formed by antagonism and competition of conspecifics, each behaviour provides a window into the evolutionary pressures that have shaped life on Earth. For the longest time, behavioural ecology studies were limited to laboriously gathered yet still incomplete observational data. The advancements of the last three decades in sensor technology, particularly solar powered accelerometers, have enabled remote collection of location and triaxial acceleration at high frequencies. The high-resolution data collected by such small, lightweight accelerometer devices opened new avenues for studying behaviour and its long-term patterns. However, analysing large amounts of complex and unstructured data is challenging. Machine - and of rapidly increasing relevance - *deep* learning have shown great promise for clustering, classifying, and modelling raw accelerometer datasets by implementing in the automated identification and classification of state and event behaviours. In this study, the recognition and classification of behaviour from accelerometer data using deep learning algorithms was examined. The focus was laid on male social behaviour events, which have often been neglected in ethograms and energy budgets. The data was collected in a wild, free-ranging population of vulturine guineafowl (*Acryllium vulturinum*) inhabiting their natural environment around Mpala Research Centre in Laikipia County, Kenya (0° N, 37° E). The vulturine guineafowl stands as an intriguing avian species with complex social behaviours that have been of interest to researchers studying multi-level societies. The accelerometers employed in this study system collect data at 20 Hz from various individuals since 2021, of which 19 were considered. To find the best possible automated approach in detecting social behaviour, different deep learning algorithms and window sizes were compared. State-of-the-art algorithms were trained on manually labelled datasets of accelerometer readings, and the best performing model was ultimately used to classify unlabelled windows across other parts of the timeseries of accelerometer data into 3 different classes, consisting of courtship, dominance, and other behaviours. The obtained frequencies and temporal patterns in social behaviour were correlated to NDVI, a proxy for vegetation status and resource availability, which are assumed to affect these behaviours. This study tried to explore the potential of deep learning for automating the study of vulturine guineafowl social behaviour in their natural habitat. This study provides a promising framework for the long-term monitoring of wild social birds and their behavioural changes over time and across changing environmental conditions in a minimally invasive and scalable way. By bridging the gap between technology and ethology, this study potentially contributes to the advancement of behavioural research methodologies and provides a foundation for further investigations into the social dynamics of vulturine guineafowl. Moreover, this methodology could be extended to the study of other social species, facilitating urgently needed cross-species comparison of the impact of climate change on behaviour.

Table of Contents

Introduction

Animal Behaviour

Animal behaviour, or ethology *(ῆθος, ethos = character*), refers to the study of everything animals do during their daily life, the corresponding underlying mental processes, and internally coordinated responses to internal and external stimuli[1]. This includes social interactions, the movement within the environment and the cognitive understanding of the surroundings[2]. Monitoring behaviour not only indicates the individual's health, welfare and productivity but can also provide information about the social interactions, population dynamics and environment[3]. Human observation, imagination and fascination with animal behaviour most probably reaches back to the very beginnings of human evolution. Intimate knowledge of an animal's habits greatly determined the success of hunting or fishing prey, escaping, or scaring away predators as well as domesticating selected species. But even outside of practical benefits, animal behaviour has all along aroused and satisfied our deep-seated interest and curiosity for the lives and minds of our pets and livestock, other real or imaginary creatures, as well as of ourselves and fellow humans[4]. As we tried to unravel the mysteries of animal minds, we gained a deeper appreciation for the rich tapestry of life that surrounds us, creating a sense of wonder and awe that transcends disciplinary boundaries. Therefore, since the dawn of civilization, humans have attributed animals with symbolic and spiritual significance, nourishing religion and philosophy alike. Delving into the historical frameworks of animal behaviour, we journey back to various ancient philosopher and thinkers, who ascribed instinct, motivation, reason, sense and feeling to animals and appealed to animal ethics, rights and even veganism[5–7]. What followed was a long period shaped by religious hierarchical worldviews positioning humans above animals and thus legitimating the dominion over the natural world. Fast forward to the 19th century, where Charles Darwin's ground-breaking work on evolution provided a framework for understanding behaviour through the lens of natural selection, sparking a revolution in scientific thought. Not only anatomical structures but also behaviours were considered adaptations existing over evolutionary times[1,2]. The emerging appreciation of the complexity and purposefulness of the actions of animals demanded long-term observations of animals in their natural settings, evolving into the fields of ecology and ethology. In the 20th century, the founders of modern ethology, Nikolaas Tinbergen and Lorenz Konrad meticulously observed, studied, and experimented on various animals in their natural surroundings, leading to deeper insights compared to impoverished laboratory environments[8]. Tinbergen claimed that the study of behaviour must address all four levels of analysis: causation, ontogeny, function and evolutionary history[8,9]. Causation explains, what makes the behaviour happen, including physiology, nervous system, hormones and cognition. The ontogeny focuses on how the behaviour develops. In other words, what developmental mechanisms lead to the occurrence of behaviour, as internal and external factors, genes, experience. The function level tries to understand how the behaviour contributes to genetic success through survival, mating, competition, or natural selection and thereby to reconstruct the evolutionary history. These four levels can help to solve the puzzle how and why individuals behave as they do[8]. These contributions still set the basics for today's research, even though the field, its tools and possibilities have changed a lot. Thanks to revolutionary technology, as GPS tracking, motion and orientation sensors, night-vision scopes, and sophisticated neuroimaging and computation, the methodologies for the study of animal behaviour have diversified. These modern applications can and should have implications for conservation and management efforts[10,11]. By deciphering the behavioural patterns of endangered species, especially in the face of global change, researchers can develop more effective strategies for their protection and preservation. Similarly, insights gleaned from studies of animal cognition and communication have far-reaching effects on animal ethics, e.g. livestock welfare, as public consciousness and perception of moral responsibilities towards other sentient beings change[6,7]. As a vibrant and interdisciplinary field, ethology continues to evolve, expand, inspire and intrigue, reminding us of the boundless wonders that await those who dare to ask.

Social Behaviour

Social behaviour refers to aggressive, mutualistic, cooperative, altruistic and parental interactions between individuals of the same species[12]. Individuals make decisions on who they interact with and how often. When individuals interact repeatedly, a social relationship between strangers, relatives or members of the same group, of same or different sex or age can develop. Sets of such relationships combine to a complex and highly dynamic social system[13,14], depending on the ever-changing connections between individuals, which can have profound effects on reproduction and survival[15]. The resulting social structure is the arrangement of relationship between individuals and groups within a society determined by patterns of behaviours and norms guiding interactions among those. The comprehension of social structure is crucial for the understanding of social dynamics, but also of the interactions between individuals and their surroundings consisting of resources, abiotic hazards, pathogens and predators, competitors, and co-operators[12,15]. The ecological and social environment determine social interactions and thus play a pivotal role in shaping population dynamics, resource distribution, and reproductive success, ultimately influencing species' survival and adaptation to their environments[12,13]. Looking at the causation and ontogeny of social behaviour, physiologically speaking they are a complex tapestry woven by (epi-) genetic[16–19], neural[20–22] and hormonal[23–25] threads, influenced by environmental factors and individual experiences[26–28]. Orchestrated by a symphony of various brain regions, neurotransmitters and hormones, social behaviours have profound effects on the social dynamics and hierarchies of a population[29,30]. Social interactions between individuals are crucial components of an animal's life and a result of (a)biotic interactions[31]. They can even be an indicator of the occurrence of extreme events such as forest fires[32] or other environmental catastrophes and problems[33], and poaching[34]. Social behaviours take many forms. Competitive or aggressive interactions determine hierarchies through dominance displays[35,36], settles territorial disputes[37], secures resources[38] and establishes social status[39–42]. Communication facilitates exchange of information between individuals through vocalizations[43], visual signals[44] and chemical cues[45–47] enabling coordination of activities, social bonding, threat displays, and danger or predator detection (alert call, cooperative defence, confusion of predator). Cooperative behaviours refer to mutually beneficial interactions such as cooperative hunting, grooming, childcare or breeding[48–50]. Courtship and mating behaviours are performed to attract mates and secure reproduction involving feeding[51], elaborate displays or ritualized movements[52,53], vocalizations and signalising fitness. Parental care comes in the forms incubation, begging and feeding, guidance to offspring, defence against predators and teaching essential skills[54]. Social bonding strengthens the social ties and promotes group cohesion trough grooming, allopreening, huddling and playing[55]. Social learning refers to the acquisition of knowledge or skills through observation, imitation or interaction with conspecifics, as parents, peers or dominant individuals within the group, e.g. foraging techniques[56,57], tools use or predator avoidance. Finally, migration or group movement stands as coordinated movement of individuals[58], ranging from synchronised flight to mass migration across landscapes, often driven by predator occurrence, seasonal changes or resource availability.

In general, ecological and environmental conditions play a key role in determining social preferences and behaviour. Under climate change, the influences of environmental factors on social interactions can become more pronounced. Changes in temperature, precipitation patterns, habitat and vegetation structure can affect the availability and distribution of resources critical to survival and reproduction as well as alter the predation pressures. Such potentially big changes in environmental conditions can trigger shifts in social behaviour (breeding phenology, competition, territorial disputes), group dynamics, dispersal and migration timing as well as foraging strategies[59–62]. Species may exhibit plasticity in their social behaviour, adjusting reproductive strategies as courtship displays and intrasexual competition or social structures in response to changing environmental conditions[63–67]. Hence, it seems evident to link obvious changes in ecological conditions to potential changes in this diverse array of (social) behaviours[64,68–70]. Currently only little is known about the social behaviour responses to changing environments, with only few studies on the impact on e.g. male courtship display[64] or aggressive interactions[69]. These research gaps in linking broadscale issues, as climate change, with behavioural adaptations can be filled with long-term monitoring, distributed across different populations and global contexts. Both within- and between-species comparison of long-term studies could enable inferences regarding the adaptations of animal under climate change[71]. Such studies contribute to our understanding of evolutionary and ecological processes, as well as future trends and could increase our ability to link genes, individual traits, behaviour and fitness with environmental variables. Looking ahead, long un-interrupted time series of social behaviour might allow the answering of questions that were not planned at the start of data collection[71]. But how is social behaviour observed and how are potential changes over time and under climate change quantified?

Observation & Quantification of Social Behaviour

Understanding animal behaviour, especially social behaviour sheds light on the adaptive strategies organisms employ to survive and thrive in a dynamic world. Each behaviour provides a window into the evolutionary pressures that have shaped life on Earth[72,73]. Studying the diversity and complexity of animal social behaviours provides insights into the evolutionary pressures shaping sociality[74]. Furthermore, social behaviour is an indicator of mental and physical states and thus of social animals' health, welfare, and subjective states[75]. However, monitoring of animal behaviour relies on direct observations, that are time consuming, labour and logistics intensive, and involve the subjective judgments of individuals[75,76]. Field observations introduce a source of limitation or bias, the observer effects on animals and their behaviour[77]. The presence of field researchers during direct observation affects animal behaviour, as they potentially perceive humans as threat or are naturally secretive or elusive[78]. Habituation to individuals or other observation units as cars is possible but labour-intensive and hence only applicable to longer-term studies, but still not guaranteeing unaffected behavioural interactions with conspecifics or non-habituated predator or competitor species[78].

Furthermore, field observations are biased by the researcher's physical limitations and proneness to give more attention to some events and individuals than others[78]. So, can we really believe what we observe?

Many social behaviours are challenging to observe, as they are rare, fine-scale behaviours, sometimes even very brief movements, so-called microevents[79]. They might also be affected by observers, not interacting socially, as if they are not disturbed. Observing and quantifying social behaviour thus presents a formidable challenge, given its nuanced and context-dependent nature, influenced by factors such as social structure, environmental context, individual differences and behavioural plasticity[80–82]. Capturing these nuances requires careful research design and data interpretation[83]. Traditional observational methods in the field, while valuable, may struggle to capture or even miss subtle, rare social interactions. This applies especially for animals which travel fast, or live in unsuitable, extreme, variable climates, unfavourable weather conditions or in inaccessible, challenging habitats or operate at night[78]. Thus, social behaviours have largely been absent from ethograms, especially those of wild, terrestrial animals[78]. However, recent technological advancements offer promising avenues for studying social behaviour with greater precision and efficiency.

Wearable Sensors

Thorough ecological ethology is based on the need to locate and observe animals to record their habits despite their potentially fast travel, challenging weather conditions, inaccessible habitats or limited night vision[78]. In the last three decades, the study of animal behaviour has more and more transformed from laborious field observation to remote observation and tracking thanks to technological advancements in and increased accessibility of wearable sensors, communication technologies and their accompanying frameworks. Thanks to their light weight, small size, low power consumption, exceptional stability, and easy integration such devices have gained in popularity[76]. This development enhanced the effectiveness of remotely tracking animal behaviours in various environments, at a larger scale than was previously achievable[76,84] and formed two major streams of objectives in the field of animal behaviour. First, the correlation of the variation in motion waveforms with energy expenditure. Second, the establishment of ethograms by inferring activity/behaviour through movement and body posture derived from such data[78].

There are several types of wearable sensors equipped with sensors for measuring motion, location, and physiological parameters, providing detailed information on animal behaviour and energetics in real-time. Their use grew rapidly in the 1990s aligning with the fast development of microprocessors and notable increases in memory capacities[85]. The kinetic characteristics (acceleration, angular velocity, etc.), pressure, and geo-location information can be accurately measured at a certain sampling rate (e.g., 10 to 100 Hz) depending on the application and identification task[76]. The power usage, battery power and memory storage affect the possible maximum sampling frequency[75].

Global Positioning Systems (GPS) have enabled researchers to track and precisely monitor individual animals in the wild, providing insights into movement patterns and habitat use over large spatial scales, facilitating studies on migration, foraging behaviour, and territoriality. Tri-axial accelerometers are the most common sensor used in animal behaviour monitoring. They are compact and low-power motion sensors that measure acceleration [m/s^2] along three perpendicular spatial axes to capture motion dynamics[3]. Due to their rapid response times and high sensitivity to movement[86], accelerometers can recognize fast changes in acceleration. They measure both a dynamic component of movement indicating the activity intensity and a static component regarding Earth's gravity indicating the posture. Doing so, accelerometers have enabled classification of many behaviours such as locomotion, resting and foraging[31,87]. This enables resolution of fine-scale animal behaviour, usually involving brief, abrupt, situation-specific manoeuvres or microevents[79]. The ability to identify those can have a big impact on the performance of the behaviour classification[79,88]. Accelerometers can be mounted to different parts of an animal, even simultaneously, which enables to expand the spectrum of well-predicted behaviours[89], and enhances the recognition performance[76,90,91]. Tri-axial gyroscopes measure orientation and angular velocity [◦/s] along three orthogonal spatial axes. They are usually integrated in with accelerometers in the same device operating at the same sampling rate to complement the captured information and can thereby improve the prediction of behaviours, which are hard to detect [76,92]. Tri-axial magnetometers detect changes in the magnetic field in particular location and measures rotation angle values (pitch, roll, yaw) [Tesla], also usually combined with accelerometer and gyroscope forming an inertial measurement unit (IMU). The simultaneous capture of linear acceleration, angular velocity, and rotation angle in IMUs has enabled a better performance in the prediction of animal behaviour[93–98]. It should be noted, that the employment of wearable sensor devices also has some impact on the animals and their behaviour[99] and that remote tracking lacks the behavioural context and demands direct observation nevertheless[78]. The combination of remote observation, which reduces observer presence effects, and direct observation keeping a high level of detail, should be the aim[78].

While the above mentioned advancements present new possibilities in observing and studying social behaviour, they create new challenges and logistical hurdles with managing such large datasets, processing the complexity and volume of the recorded data and with performing high-throughput behavioural data analysis[85].

Analysis of Sensor Data using Machine & Deep Learning

To make inferences about daily life activities or energy expenditures from sensor data, an elaborate analysis technique is required[76]. The processing is usually done after data collection, but collecting and storing raw sensor data for later processing is inefficient and unscalable[3]. Downloading the raw data via antenna is also not very advantageous[3]. To analyse such complex signals will require the development of efficient but still accurate methods[79]. Already existing methods are split into two categories, semi-automated and automated approaches.

Semiautomated AAR involves the manual characterisation of the sensor signal patterns followed by a classification using a decision tree[100]. This approach is not limited by a fixed-size sliding window but requires appreciable investment in time and understanding[79,100].
Automated animal activity recognition (AAR) enables the monitoring of the variability in animal behaviours across time. With huge improvements in sensor technologies and computation, exellent successes in AAR have been achieved[76]. In automated AAR the sensor data is first segmented into windows of fixed size. Then, either the raw, segmented data is fed to a classifier or descriptive features are derived from that window[31]. These features are usually statistical summaries of the sensor data, such as mean, standard deviation, skewness, vector of dynamic body acceleration (VeDBA), overall dynamic body acceleration (ODBA)[101]. The extraction and selection of such features heavily relies on human expertise and a very precise and adapted pre-processing[85]. The defined window size for the feature computation depends on the identification task, especially on the frequency/duration of the focal behaviours[79,88]. The behaviours are then separated into clusters or classes using simple thresholds[102,103] or machine learning algorithms[87,104–106]. If the collected time-series of sensor data is kept untouched and raw, usually the data is fed into diverse array of machine and deep learning models to classify the data into behavioural categories[76].

Machine learning, as a very promising data processing and analysis technique, has been widely applied to animal behavioural classification based on data collected by wearable sensors[76]. Such modelling methods include linear regressions, support-vector machines, decision trees, linear/quadratic discriminant analysis, and random forest approaches[76]. Generally, to accurately classify animal behaviours through these methods, manual feature extraction and selection are required. But these processes are time consuming and heavily rely on expert knowledge, which leads to feature extraction and selection challenges[107]. There are some approaches available which combine feature extraction with feature selection. Thus, the most discriminative features for the specific classification task is automatically captured, while also providing interpretability and insights by ranking the importance of different features for predicting the target[76].

Deep learning, as a more recent branch of machine learning, has been showing an excellent automated/integrated feature-extraction ability while usually requiring less pre-processing than traditional methods[76,108,109]. Deep learning models combined with wearable sensors have revealed promising performance in distinguishing daily animal activities[75,76,90,98,110] using different classification algorithms. For example, Feed-forward Neural Networks have been used on large datasets, Convolutional Neural Networks on image and time series classification, Recurrent Neural Networks, especially Long Short-Term Memory on sequential data and time series, as well as hybrids models and Autoencoders.

Usual challenges during the development of deep-learning models for AAR, include annotation scarcity, class imbalance, inter-activity similarity, energy efficiency, multimodal fusion, domain generalization, and open-set recognition. To solve these challenges, dedicated deep-learning models are required[76].
Even after laborious, time-consuming labelling annotation scarcity can still occur and often results in an overfitting model and poor generalization performance, limiting the applicability of models to real-world AAR scenarios. Data augmentation is a low-cost pre-processing technique to create new samples through the transformation of existing annotated data via various approaches to

expand data size and thus promote classification performance of deep learning models[76,111–113]. Another approach is semi-supervised learning where unlabelled data is used to assign pseudo-labels[114].

Class imbalance occurs where the frequencies are inconsistent across different behaviours. Annotating rare or infrequent behaviours is difficult because they occur occasionally or for short durations[76,109]. Deep learning methods trained on imbalanced datasets are usually biased towards the majority classes. This causes a decreased model generalizability and higher misclassification rates for the under-represented categories[76]. To overcome this limitation there are different techniques, as resampling with either over-sampling the minority class or under-sampling the majority class to balance the class distribution[115]. In scenarios of extremely imbalanced datasets, the classification tasks can be reformulated as an anomaly detection problem, in which minority-class instances that are dissimilar to the majority class are treated as outliers or anomalies[116]. In this case, one-class classification methods can be used to build a model that learns on the majority class characteristics and then distinguishes them from the minority class[117].

Inter-activity similarity occurs when different animal activities have similar characteristics or movement patterns[95,109,118]. This affects the ability of deep learning models to extract distinguishable features that uniquely represent behaviours, leading to high confusion in the classification/class prediction[119]. Active and inactive behaviour very easy to distinguish, but within those categories it can become tricky[78]. Employing a fine-grained activity recognition, seeking to recognize subtle differences between similar activities by using more such detailed features can provide remedy to this issue[76]. One option is the combination of sensors (GPS, IMUs, heart rate logger, …) to better distinguish behaviours based on other parameters. This can possibly lead to very new and changing insights into an animal's life[78]. This combination of multiple wearable sensors, a so-called multimodal fusion, helps to receive richer information to better distinguish behaviours. Sometimes, sensors of different types are mounted to an animal to record diverse characteristics. Combining these sensors tends to result in improved performance in animal behaviour classification tasks compared with using only one modality[3,90,94,107] but the models may struggle to generalize[110], as conflicting correlations between multiple modalities can result in limited recognition performance[107]. For social interactions particularly, a precise spatial proximity tracker can support the recognition and classification[78]. Another approach is context-aware modelling, where e.g. the time of day, the location or environmental conditions are included to enable effective clarification of the purpose of the activity[76,120].

An additional issue that can occur is the open-set recognition problem. Most training datasets only cover a part of the full spectrum of the specific animal activities. Thus, some rare or infrequent activities, which are nonetheless important and occur in real-world monitoring scenarios, can be absent from training datasets. Consequently, these unseen behaviours are often misclassified into known behaviour categories in a training dataset. A holistic model does not only accurately classify known categories but also effectively deals with unknown behavioural categories[121].

Sensor-based data from animals, analysed with deep learning models, has already found many applications, especially in livestock or animals in captivity. There are many studies on mammals (mainly cattle and pinnipeds) with 45% of the studied species, birds with 34% (mainly penguins and seabirds), fish 11% (mainly sharks), a few reptiles and very few other taxa (cuttlefish, squid, toads)[78]. In smart farming, it is applied to estimate growth or monitor disease. There are only few reports on animal behaviour classification, but they have typically been limited to either traditional machine learning or specific animals and behaviours[76,89,110].
Most of the studied animals are livestock as part of smart farming, with a big focus on the detection of disease or lameness, oestrus or onset of calving. Studies in sheep, horses chicken and dogs have

exclusively focused on state behaviours based on economic losses[75,76,90,93,109–111,122–124]. Social interactions have only been identified in cattle and pigs. In cattle, as ruminants with economic value as well, there have been studies on state behaviours but also social behaviours, as social licking, headbutting, attacking or mating, with various subsequent ecological questions[76,94,96,98]. Pigs have been subject to studies to improve health monitoring and understand their nursing, breeding, parturition but also playing behaviour[113,125,126]. But research gaps remain in the identification of social behaviours from wild, free-ranging, terrestrial animals, with only very few studies(SOURCE)[78].

Problem Statement

The rise of accelerometery has been helping to circumvent the age-old limits of direct observation. By using accelerometers, the movement behaviour of wild animals can be measured during important events and periods, while being basically unlimited by visibility, observer bias, or geographic scale[78]. The combination of such sensor data and deep learning analysis tools facilitates the development of systems capable of accurately detecting, classifying and monitoring various activities/behaviours, potentially revolutionizing research, animal management and promoting animal health and welfare[76,110]. As collecting and storing, then transferring wirelessly and finally processing the sensor data is very inefficient, unscalable and disadvantageous, there have lately been some efforts to establish a real-time in-situ behaviour classification on embedded systems of the sensor device, only needing to store the predicted behavioural class[3].

Analyses of sensor data to identify social behaviours are very few in number compared to state behaviours especially of wild terrestrial species, compared to wild aquatic, domestic or captive species. Social behaviours of wild terrestrial species have largely been absent from ethograms[78], mainly because of their fine-scale nature of social behaviours as they often are short, impulsive movements. The identification of such microevents using existing models, still remains poorly studied but could help to exploit the full potential of acceleration data in animal behaviour classification. In general, substantial challenges remain in getting the most out of accelerometers, because of the management, validation, calibration and analysis of such big data[78]. Deep learning algorithms has been only poorly exploited in movement ecology[85].

Vulturine guineafowls (*Acyrillium vulturinum*) are a small-brained bird species that are native to East Africa, particularly living in savannas, scrublands and dry woodlands, foraging on seeds, fruits, insects and small invertebrates. They are highly social, cooperatively breeding bird forming groups with up to 65 individuals, with their home ranges overlapping in time and space[127]. Thereby they form multi-level societies which are highly hierarchical with much male dominance display, like chasing each other, occurring. Vulturine guineafowls start breeding during rainy season when vegetation cover and food availability is higher. A preceding period of male courtship display, especially bowing, and intra-sexual competition leads to the formation of mating pairs[54,127]. Even though vulturine guineafowls have a high habitat fidelity, it remains very difficult to observe them over a long period of time, as they disappear into the inaccessible parts of the savannah, like dense bushes, for the hotter parts of the day.

So, the aim of this study was to investigate the ability of deep learning algorithms to identify vulturine guineafowl male courtship and dominance behaviours from labelled accelerometer data, which have not been included in previous ethograms. The performances of different algorithms were compared in a two identification tasks. Furthermore, the best performing model was used to predict unlabelled parts of the accelerometer timeseries. These obtained behavioural frequencies were then correlated with the normalized difference vegetation index (NDVI), to evaluate a potential predictor of these social behaviours. It was hypothesized that the deep learning models will be able to recognize the courtship and dominance behaviours and the derived frequencies can be explained by the NDVI, as a proxy of vegetation and resource availability[128].

To do so, field-borne video recordings were used to annotate accelerometer readings, deployed on a free-ranging population of vulturine guineafowl around the Mpala Research Centre in Laikipia County, Kenya, with behavioural categories. The labelled behavioural data was then fed to various deep learning algorithms, learning to accurately identify and classify social behaviours. Later a well-performing, trained algorithm was used to predict other parts of the recorded accelerometer data without ground truth labels. The hereby quantified frequencies over time were used to monitor behavioural changes and answer ecological questions associated with reproduction and competition over the course of the years 2022 and 2023. This study tried to establish a protocol of best practices for data acquisition and analysis for future studies being able to include a longer time-series of data. Drought is the one of the most challenging aspects of climate change in East Africa, as precipitation and water storage decline especially during rainy seasons[129]. Future studies could include the monitoring of courtship and dominance behaviours in the face of climate change, as they are crucial for the survival of a not yet endangered species.

Methodology

Study System

This research was conducted under the Vulturine Guineafowl Project, which was established in 2016 and is currently funded by an ERC Starting Grant until 2024. The project's base is located at the Mpala Research Centre in Laikipia County, Kenya.

Study Site

The Mpala Research Centre is located within the Mpala Research Conservancy (0°17'31" N, 36.53'54" E), just above the equator around 40 km North of Nanyuki, the capital of Laikipia county. The conservancy is part of the upper Ewaso River Basin, on an elevation of 1600 meters above sea level.

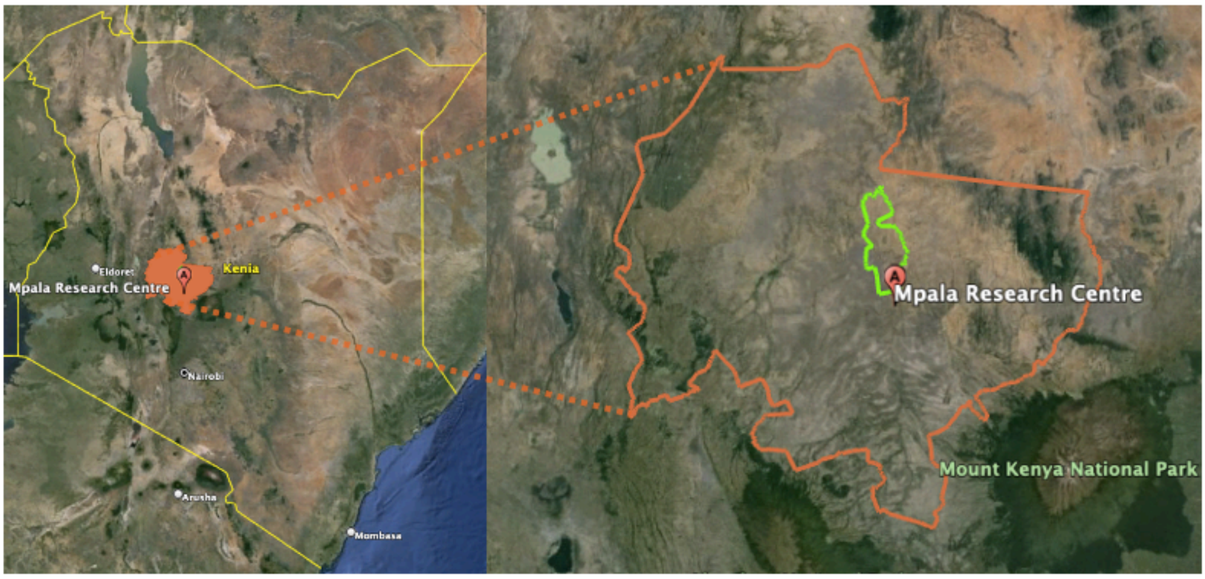


Figure 1: Study site at Mpala Research Centre within Mpala Research Conservancy (light-green) and Laikipia County (orange) in Kenya (Credits: Google Earth Pro)

Climate, Soils and Vegetation

The yearly mean temperature is approximately 17°C. The conservancy is located North-West of Mount Kenya and experiences around 400 mm of total precipitation a year[130]. As the field site is very close to the equator, the seasonal climate is characterised by two wet seasons with higher rainfall, around April-May (long) and October-November (short), and two dry seasons with low rainfall[127,131,132]. The shorter wet season is followed by a longer period of drier conditions, usually reaching from December to February, but in some years extending into April causing drought. The long rains occur starting from March, sometimes lasting until June. Each wet season is usually followed by an intermediate season, where vegetation remains lush with occasional rainfalls. If the total precipitations remains low for a longer period, conditions become extremely dry with trees losing their leaves and the short vegetation dying off. The intensity and duration of the rainfall periods determine the vegetation cover and the availability of insects, an important part of the diet of vulturine guineafowl besides grass roots, seeds, and other small invertebrates[127].

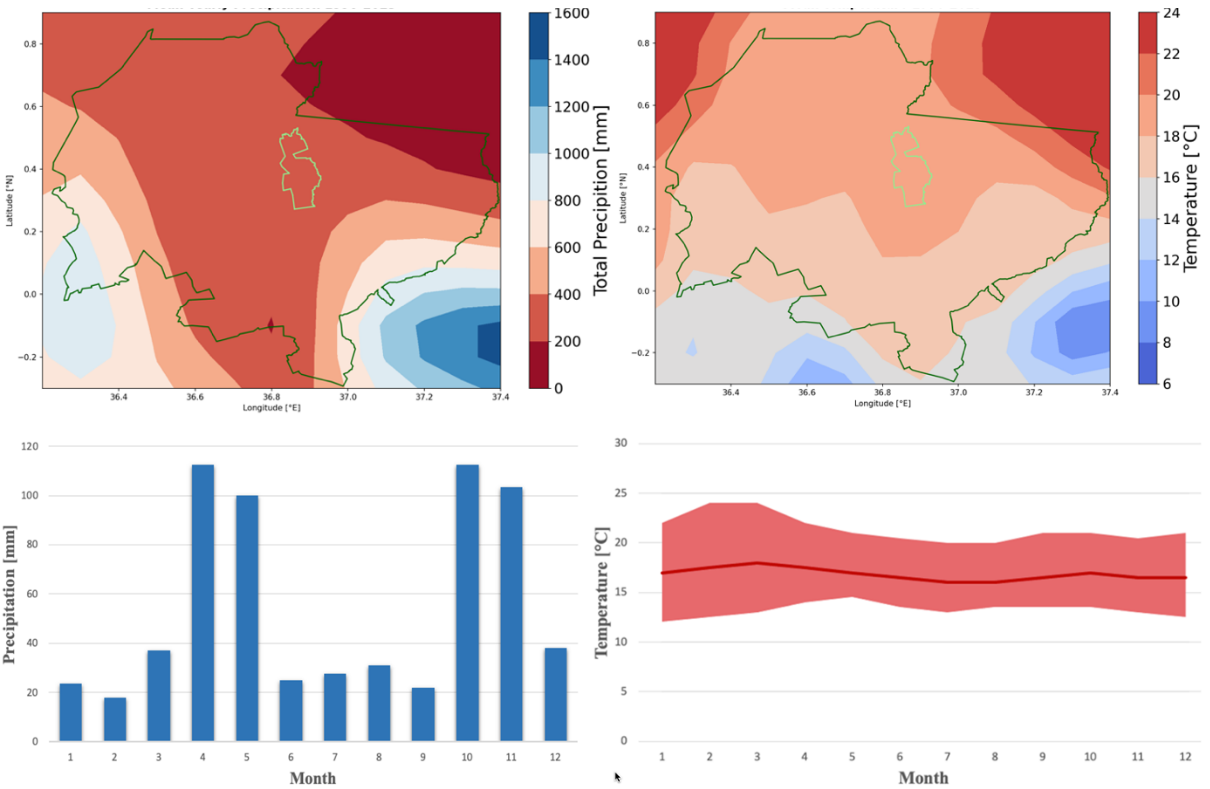


Figure 2: Map of total yearly precipitation for the reference period 1990-2023 (top left). Map of mean temperature for the reference period 1990-2023 (top right). The mean monthly total precipitation (bottom left) and mean monthly temperature (bottom right). All data derived from the ERA5-Land monthly averaged data[133].

The Mpala Research Conservancy is characterized by a semi-arid savanna habitat, with five main soil types which are mainly covered with Acacia bush- and scrubland, Acacia thicket, dwarf bush grassland and grassland[130]. The vulturine guineafowls have predominantly specialized on the red Luvisols which are dominated by *Acacia mellifera* and *Acacia etbaica*[48].

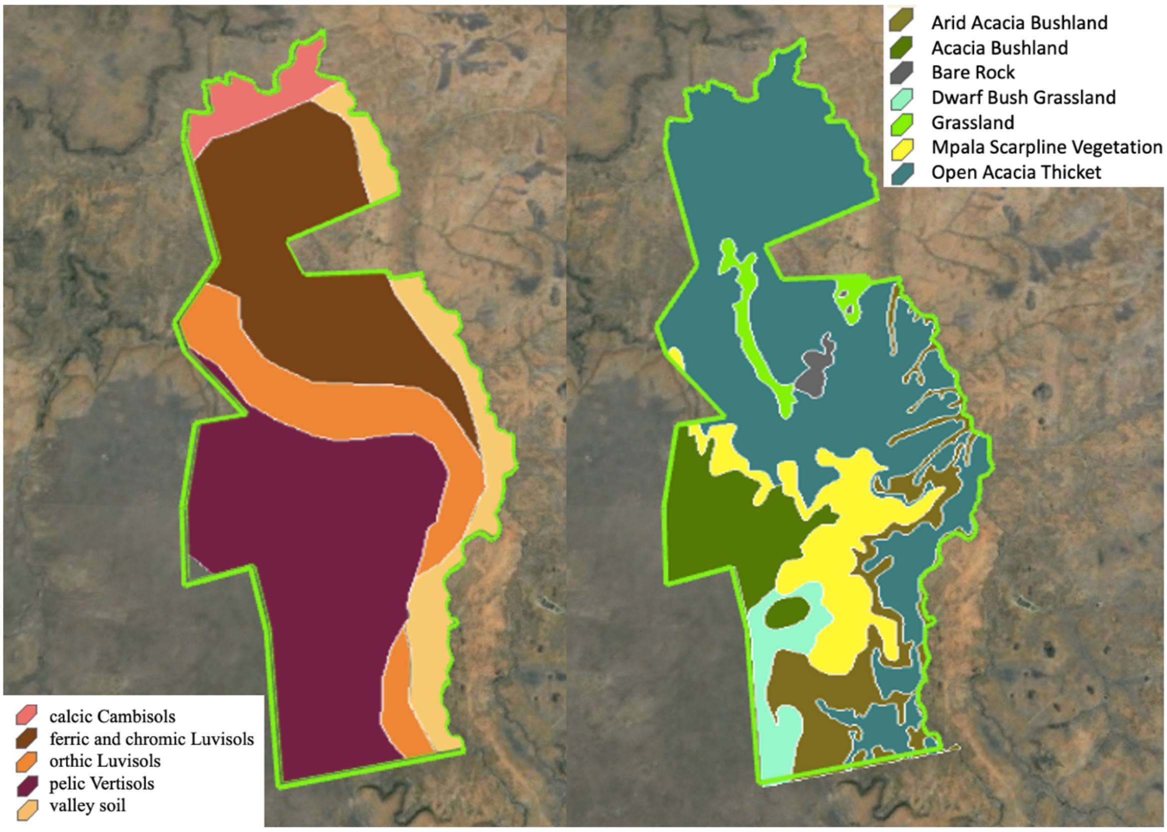


Figure 3: Soil and vegetation types across Mpala Research Conservancy, derived from shape-files which were provided by John Gitonga, research assistant at Mpala Research Centre (Credits background map: Google Earth Pro)

Study Species & Population

This study was conducted on a free-ranging population of vulturine guineafowl (*Acryllium vulturinum*) living in proximity of the Mpala Research Centre. The population is habituated to cars allowing their observation in areas accessible for cars. The vulturine guineafowl is a mainly terrestrial, small-brained and group-living bird forming multi-level societies[131]. They live in large, multi-male, multi-female groups, reaching from 13 to 65 individuals, which remain stable over time. The groups' home ranges overlap in space and time[131]. These social groups contain many adults, sub-adults and juveniles[131].They are a social bird species, showing male dominance displays, a complex courtship behaviour and cooperative breeding[48,127], which occurs during elongated periods of rain. Even though they have a small brain to body ratio, the vulturine guineafowl is able to maintain many different social relations across time and space, challenging the conception that multilevel societies are exclusive to large-brained mammals[131]. Vulturine guineafowls show significant differences in behaviour and collective movement patterns between wet and dry seasons and the corresponding variation in environmental conditions. Social behaviour, locomotion, home range overlap and roosting are dependent on dry and wet seasons[131]. These behavioural adaptation strategies are most likely crucial to dampen the impacts of climate change on the individual fitness[127,134,135]. Such strong seasonality leads to a dynamic reorganization of the social system, where species switch from being territorial to group-living also triggering a shift in the group structures, so-called fission-fusion dynamics[136,137]. Increasing drought under climate change could dramatically disturb these. But in environments with strong seasonality, natural plasticity in social organisation and movement decisions on foraging trips and dispersal could dampen the effects of environmental variations on physiology[127].

Social Behaviours

The vulturine guineafowl is a gregarious species with steep dominance hierarchies. The high-cost aggressive interactions towards males with ranks closely around their own are strategically performed. These dominance interactions are used for access to resources, but they are costly and hence should be strategically deployed, most efficiently targeted at the individuals closest in rank for most gain[138]. Directing chasing interactions towards close competitors can stabilize hierarchy[138]. Chases are the highest-cost interaction, as they deplete energy reserves, include risk of injury and predation, but these costs usually associated with dominance interactions may not be so relevant in vulturine guineafowl. These aggressive interactions are primarily directed towards males one to three ranks below themselves, according to the close competitor strategy, which is observed in many species. These aggressive interactions form the dominance hierarchies which have consequences on access to food, roosting positions and reproductive opportunities. Male vulturine guineafowl usually engage in dominance interactions most frequently, probably caused by the dominance of males over females in the group[138].

The dominance behaviour is characterised by erupting and chasing another male away, often preceded and followed by a distinctive posture. This posture is characterised by standing on tip toes, stretching, and bending the whole body, especially the neck into the air. This behaviour is referred to as CHA.

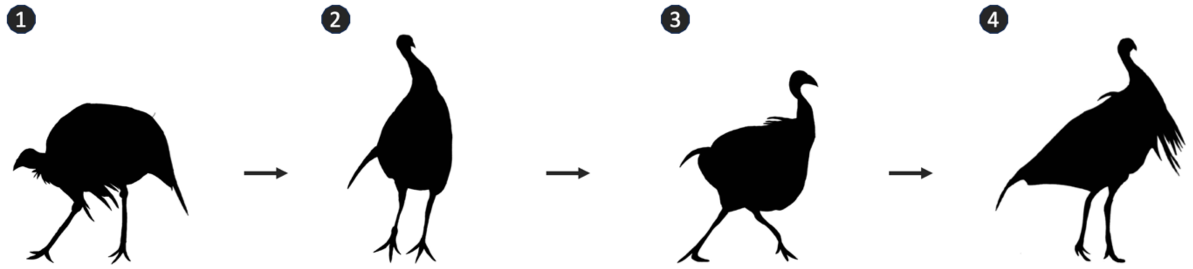


Figure 4: Chasing is mostly initiated from foraging behaviour (1). It starts with a tiptoe posture (2), followed by chasing another individual away (3) and ends with another dominance posture (4).

Vulturine guineafowls invest in reproduction during seasons with high resource availability,. During scarce periods they invest in survival[127]. Vulturine guineafowls live in large stable groups for biggest part of the year. but then at the beginning of the wet and hence breeding seasons, following a longer period of intense courtship displays, they start forming pairs[127]. The pairs move separately from the rest of the group with the male mate-guarding the female. As a ground nesting bird, the females then lay and independently incubate 7 to 12 eggs in a scrape on the ground with high predation risk[139]. Vulturine guineafowls possibly are both plural and cooperative breeders[48]. The babysitting and chick guarding is cooperatively distributed among group members, while the chicks maintain a close adult relation. The offered help is significantly male-biased. This cooperative breeding allows the female to recover from her natal investment. The chicks are very vulnerable to predation during first weeks of their lives. Therefore they benefit from protection by within-group helpers. The extremely sex-biased dispersal with all males staying in their natal groups[140], causing a high relatedness among males within the group, could explain the indirect fitness benefits of cooperative breeding[141].

The courtship behaviour, called bowing, is a male display, in which, when protein-rich food is found during foraging, e.g. insects, worms or larvae, the male runs a few meters, drops the bit of food and then presents his findings with another distinctive posture bending the neck, awaiting a female to react and potentially allowing a mating relationship to build up. This behaviour will be referred to as BOW.

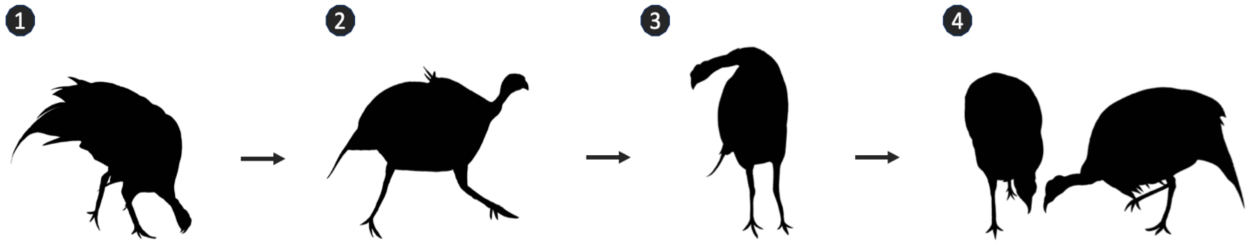


Figure 5: Bowing starts with finding a special bit of food (1). Then, the male runs a few meters (2), drops the food and presents it with a distinct stretching from tiptoe to neck posture (3). Best case scenario: female individual approaches fast and thankfully picks up the food (4).

Research Design

During this study, a best practice approach for animal activity recognition according to Mao et al., 2023 was applied. The first goal was to increase the number of video-recorded social behaviours from vulturine guineafowls tagged with an accelerometer in the field. These video-recordings were used to annotate and create labels serving as ground truth for the algorithm training. This process included equipping and video-recording as many individuals as possible to include as much variability as possible over as many periods as possible. The meticulous annotation phase was followed by a thorough data quality control, as data gaps and imprecisions in the annotation software occurred quite frequently, unfortunately leading to a loss of labelled social behaviours. Then, the labelled sequences were merged with the Movebank database of accelerometer recordings via exact timestamps. The data subsequently was segmented into different window sizes with a 50% overlap. This thorough pre-processing phase created the best possible datasets to train the different deep learning models. The subsequent trial-and-error based phase of classifier training was followed by the analysis of the model performances. The best performing model was ultimately used to predict unlabelled accelerometer readings to quantify the distribution and frequency of the focal behaviours over time and seasons and were correlated with a timeseries of Normalized Difference Vegetation Index (NDVI), as a proxy for vegetation status.

Identification tasks

The first identification task was to compare the performance of different algorithms in distinguishing the two social behaviour classes from other behaviours assembled in the NOT class. A second identification task was the comparison of different window sizes feeding the algorithms. A commonly applied random split was applied to split into training-validation and test sets. Then a 10-fold cross-validation procedure, stratified across the behavioural classes, was applied. There are other approaches to better investigate the generalization ability, as individual-based splitting or time-stratified splitting. But these methods were neglected in this study due to the small number of annotated social behaviours, which were also highly imbalanced across birds and time. Furthermore, several studies reported significant decreases in performance with these stratified splitting methods[142]. The individual-based splitting method, or Leave-one/some-individual(s)-out (LOIO/LSIO), investigates the inter-individual variability and how well the trained model performs on unseen individuals potentially unravelling inter-individual differences in their behavioural patterns[143]. The time-based splitting method, or Leave-one/some-time-period(s)-out (LOTO/LSTO) analyses the variability across time and how well the trained model performs on unseen time periods, potentially illustrating changes in behaviour or tag orientation over time[89,144]. These methods were not applied here, but should be included in further studies, as soon as more labelled social behaviours are available. As these tasks were not applied, the final prediction of unlabelled data was only employed on known individuals, for which social behaviour labels existed, and on known time periods, for which social behaviour labels existed. This was assumed to maintain the error rate small.

Data Collection & Sampling Methods

The first birds were banded at the beginning of the project in 2016. Since then various individuals have been equipped with accelerometers from e-obs GmbH[145] simultaneously and continuously tracking acceleration at 20 Hz and GPS at 1 Hz. The acceleration is the rate of change of velocity and is measured in m/s^2 along three perpendicular axes called X, Y and Z. It is influenced by the device's orientation and accelerated movement of the device[145]. Some of the tags not only include a GPS and an accelerometer, but also a gyroscope, allowing a detailed visualization of the absolute orientation in the three-dimensional space. The additional modules of the so-called Inertial Measurement Unit (IMU) could be brought in for future studies to better distinguish behaviours. The accelerometers tags contain an accumulator cell fed by solar panels on the top of the tag, potentially creating issues with cloudy weather conditions, dirt and dust, feathers or wings obscuring the panels or the tag moving into unfavourable positions. They can transmit the data remotely via antennas to the e-obs BaseStation, which must be inside the detectable range.

To equip new birds with accelerometer tags, the birds must be trapped, but they must become habituated to a trapping set up. A trap is set up step by step over the course of a few days. During this process the birds are baited into the trap every morning, so that a maximal number of birds enters the trap after a few days, without getting suspicious. Then, on due day, the trap is connected to a remote control, capable of triggering the front of the net to fall down. After trapping the birds, they are put into dark cages into the shade, so that stress level is held low and they do not overheat during the ringing and tagging process.

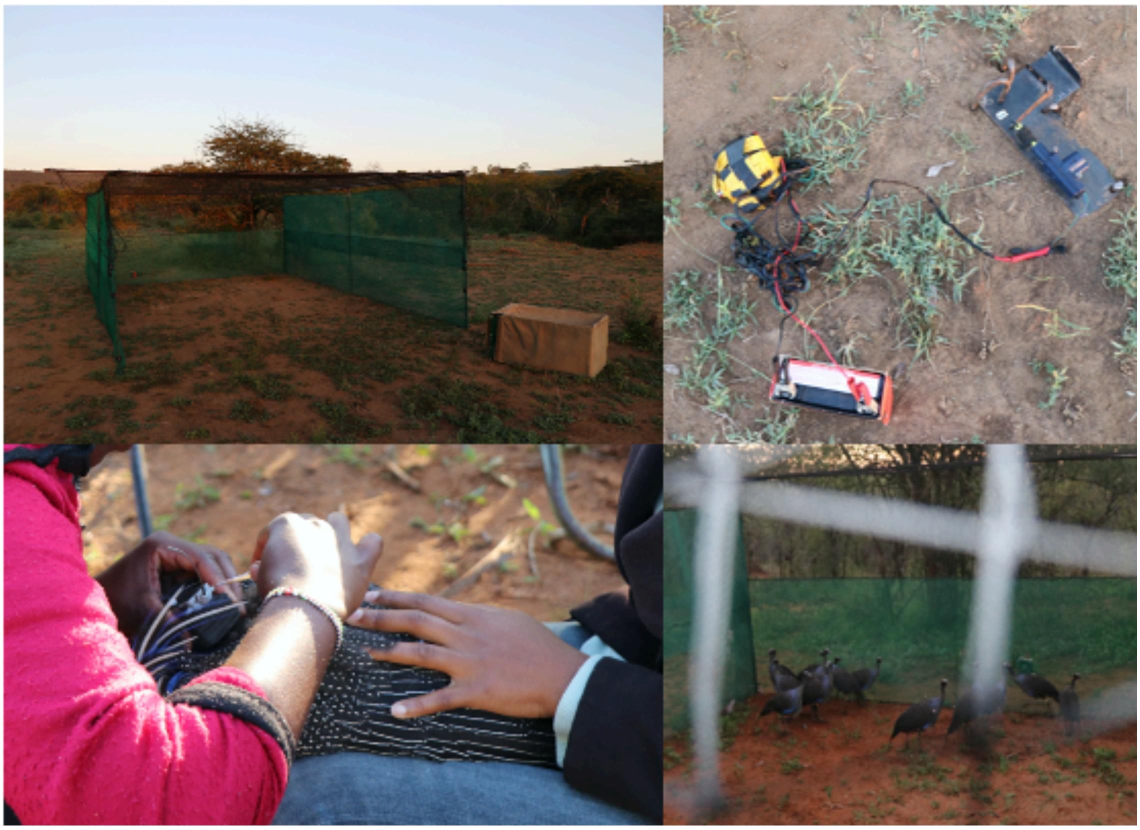


Figure 6: The trap is set up step by step (top left). On due day the net above the entry is triggered via remote control (top right). After trapping, they are weighed, measured, ringed and tagged if necessary (bottom)

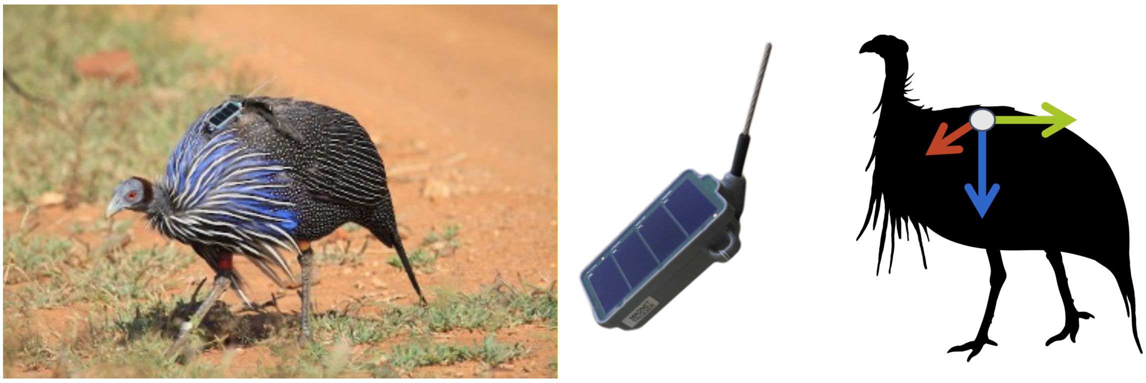


Figure 7: A tagged vulturine guineafowl with the e-obs device. The acceleration is measured along the three illustrated axes

The considered accelerometer recordings have been obtained from 19 different individuals belonging to 5 different social groups recording since different trapping events. Most of the accelerometers record only in the mornings from 6:30 AM to as late as 11:00 AM, as recording for longer periods would use up too much battery and storage on the integrated memory. The data must be downloaded regularly, best every night, so that the memory is not used to full capacity. The downloaded GPS and accelerometer data is then uploaded onto the long-term database on Movebank[146], a global repository for animal movement data, which are publicly available on movebank.org. The acceleration data can be visualized for quality control in the open-source tool Movebank Acceleration Viewer[147].

Annotation & Quality Control

Annotating animal behaviour is generally arduous and challenging, A total of over six hours (06h 13min 22s) of video material were considered to create the behavioural labels. The video material was recorded during field seasons spread across the years 2021 to 2023. Around 50 percent was recorded in the years 2021 and 2022 by other members of the research group, the other 50 percent produced during this study's field work season from April to June 2023. The birds were recorded out of the cars, which they are habituated to, using the video camera. The birds were often found one of the several accessible glades, especially during the first two to three hours of sunlight, as they disappear into the bushes with increasing solar radiation and temperature. First a GPS time clock was filmed on a smartphone to later synchronize the video material exactly. Then, the rolling camera was focused on tagged individuals, hoping to capture a social behaviour, as bowing or chasing. It is not unusual to overlook some events of such rare behaviours, even for domain experts as the long-term field team, as they happen occasionally and in very short durations and often outside the observer's focal range of view. During the field season 2023, one could observe an increasing frequency of the two focal behaviours, especially bowing, as the rains started mid of April, which transformed the very dry savannah into a greener landscape. The video material was later sighted and annotated after the observed bird's ID and datetime. In the sound and image annotation program ELAN (version 6.4)[148], the videos were time synchronized and labelled creating an output txt-file with exact start and end timestamps for each behaviour. The output of ELAN was only considering a 10th of a second. To better capture these short social interactions, more precision was necessary, focusing down to milliseconds. Thus, the data was visualized in the Movebank Acceleration Viewer[147] to check the quality and true synchronicity of the start and end timestamps with the acceleration measurements. The exact beginning and ending of each focal behaviour was adapted in the txt-file. The produced number of labels is illustrated in table 1.

birdID	tag_type	sex	# BOW samples	# CHA samples	x̄ Duration BOW [s]	x̄ Duration CHA [s]
W1744	ACC	M	0	0	NA	NA
W1393	ACC	M	0	4	NA	0.87
W1732	ACC	M	0	6	NA	0.74
W1520	ACC	M	1	0	1.91	NA
W1415	IMU	M	0	1	NA	0.63
WT00500	IMU	F	0	3	NA	1.61
WT00162	IMU	M	3	3	1.86	0.89
W1413	IMU	M	2	11	2.09	0.92
W1307	ACC	M	4	4	2.06	1.62
WT00043	ACC	M	4	9	1.50	0.83
WT000625	ACC	M	1	1	2.61	0.74
W2625	ACC	F	0	1	NA	2.45
W1309	ACC	F	0	0	NA	NA
WT00584	ACC	M	0	3	NA	0.98
W1501	ACC	M	1	0	1.99	NA
W1686	ACC	M	1	0	1.71	NA
WT00044	ACC	M	9	3	2.15	1.06
WT00580	ACC	M	13	0	1.97	NA
W1430	ACC	M	5	4	3.56	0.87
			44	**53**	**2.13**	**1.09**

Table 1: The individuals, their tag types and sex, as well as the recorded BOW and CHA events and the duration of their behavioural expression. The marked individuals in grey were used to predict unlabelled data from.

Pre-Processing

The txt-files with the starting and end times were loaded into RStudio (version 2023.09.0+463) to merge the timestamps with the accelerometer data stored on Movebank using the "move" (version 4.24)[149] package. A data frame with X, Y and Z axes readings every 0.05 seconds for all the video-recorded birds and periods was obtained. The acceleration readings were standard normalized, as it improves the classification performance and increases the training speed[3,76]. Previous feature extraction was not needed, as deep learning is doing that automatically[76,108,109]. The normalized accelerometer readings were either processed and fed to the algorithms as a raw dataset or as images of the acceleration graphs for two different window sizes (1 or 3 seconds) with a 50% overlap respectively. All created windows were also labelled with their distinctive position in a consecutive sequence of behaviours to feed them to the algorithms including an LSTM, as the CNN-LSTM and LSTM-CNN, as they require sequences of time series data. The raw datasets were fed to the MLP and TCDA-CNN classifiers, while the images were fed to the CNN, Autoencoder, ResNet, CNN-LSTM and LSTM-CNN. The images were created by plotting the accelerometer axes in red, green and blue on a white background using the ggplot[150] package and then were saved as RGB images of shape (64,128,3). The window segments consisted of 20 or 60 consecutive acceleration readings for the 1 or 3 seconds respectively, as it was recorded at 20 Hz. The window size impacts the performance of the models in that some behaviours, of different length, are better characterized than others[151–153]. To determine the optimal windows size, these two window sizes were compared after model training. The segmentation process generated windows that sometimes contain the expression of several behaviours. Such windows were labelled according to the majority rule of readings. This method created a small number of segments labelled with social behaviours and a much larger number for the NOT class (table 2).

		BOW	CHA	NOT
50% overlap	**1second**	296	240	19172
	3seconds	137	128	12235

Table 2: Numbers of labelled segments for each window size and behavioural class.

Establishing animal behaviour classification models that perform well on rare behaviour, especially on the unseen test set, is challenging. But the resulting class imbalance is not the main cause of the lower metrics scores for these rare behaviours. The scarcity of training data for these behavioural categories is responsible for that[3,154–156]. There are various methods for balancing the datasets, such as under-sampling, over-sampling, weighting the datapoints by the inverse of the frequency of their corresponding class, and creating new datapoints. But these methods can not get rid of the issues related to annotation scarcity completely, hence finding well performing approaches for rare behaviours remains a subject of ongoing research[3]. In this case, the dataset was separated into training, validation and test set retaining a close to natural ratio of BOW : CHA : NOT (15:15:70) by under-sampling the NOT class to 5000 samples and simultaneously augmenting the under-represented classes by applying class weights during the classifier training.

Cross-Validation

After a test set (test set percentage = 0.15) was randomly held out for the final evaluation, a 10-fold cross-validation stratified across behavioural classes was applied to not further decrease the number of samples which available for learning the model as compared to define a fixed validation set. Another argument was that the outcome of the model strongly depends on a particular random choice of splitting into training, validation and test sets[157]. Thus, the stratified 10-fold cross-validation was applied across all model training to monitor the predictive performance over different subsets of training data. In this basic approach the training set is split into k smaller sets maintaining a stratification across the classes. The model is trained using k−1 subsets as training data and the resulting model is validated on the left-out part of the data. This method is computationally more expensive and creates longer processing times, but it does not waste much data. This method offers the major advantage in problems where the number of samples is very small[157].

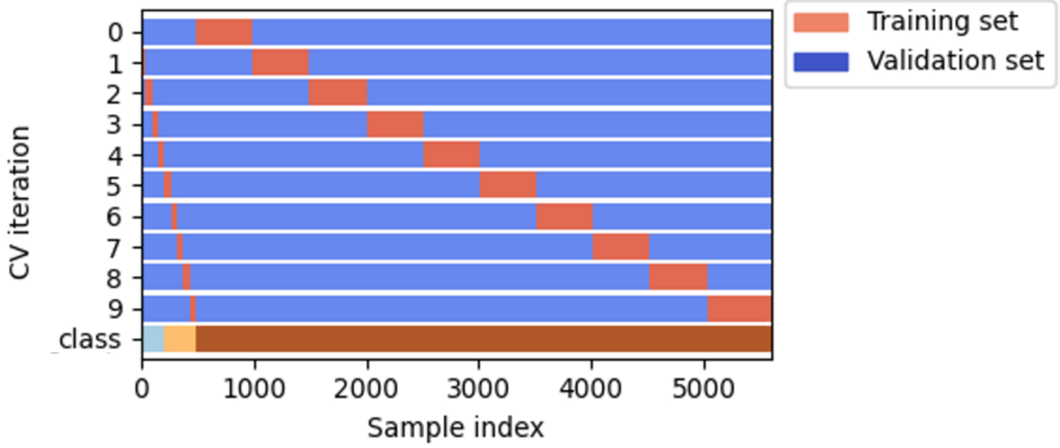


Figure 8: Visualization of the stratified 10-fold cross-validation

Algorithms

The analysed algorithms included many different types of neural networks. Neural networks are able to automatically detect by themselves very complex and highly discriminating features and patterns in the data[85,158]. A Convolutional Neural Network (CNN), a Triple Cross-Domain Attention CNN (TCDA-CNN), an Autoencoder, a Residual Neural Network (ResNet), a CNN combined with a Long Short-Term Memory (CNN-LSTM), a LSTM-CNN and finally a Multi-layer Perceptron (MLP) classifier, serving as comparison benchmark, were used. The comparison with the baseline classifier investigates if the increased complexity in the other neural networks really leads to better performances, as the MLP is still supposed to generate better outcomes than random choice. The models' hyperparameters were mainly tuned on a trial-and-error basis, but also applying a grid search for some parameters using the scikit-learn library (version 1.4.1)[159]. The models were implemented in Python's "TensorFlow" (version 2.15.0)[160] and "Keras" (version 3.0)[161].

Multi-Layer Perceptron (MLP)

An MLP was used as a baseline classifier. The MLP is a Feed-Forward Neural Network, or perceptron, which have their roots in the human brain research, as it was intended to understand how visual data is processed and how objects are recognized by the brain[108]. This network type has found applications in AAR in several studies, as they can handle large datasets and perform very well despite their low complexity[3,93,94,154,162,163]. Their main components are neurons (or units), that take the input data or the output from the preceding neuron and produce a single output. This output is input to every neuron in the next layer. From the connections and weights of each layer's outputs, the activation function computes the input into the next layer. In the last layer, the output is generated by applying the sigmoid (binary classification) or softmax (multi-class) activation function. Each network layer only takes information from preceding layer and only gives it to subsequent layer, hence feed-forward, allowing it to learn complex non-linear relationships between inputs and outputs[108,164]. The applied architecture for the MLP included an input layer, three hidden, fully-connected dense layers with 128, 256 and 128 neurons respectively and finally an output dense layer with three units equal to the number of classes. The activation functions were ReLu for the dense layers, and softmax for the output layer. The applied optimizer was the Adaptive Moment Estimation (Adam) with a learning rate schedule and the loss function Categorical Cross-Entropy. The model was trained over 100 epochs using a 10-fold cross-validation, so 10 epochs per fold.

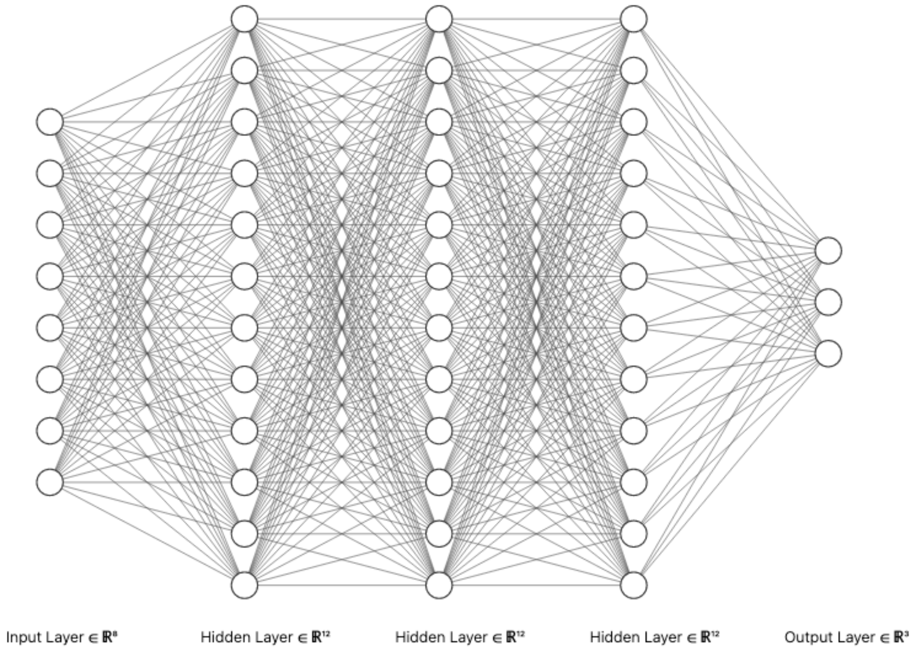

Input Layer $\in \mathbb{R}^8$
Hidden Layer $\in \mathbb{R}^{12}$
Hidden Layer $\in \mathbb{R}^{12}$
Hidden Layer $\in \mathbb{R}^{12}$
Output Layer $\in \mathbb{R}^3$

Figure 9: Dot-Visualization of an MLP, where each dot represents a neuron and the arrows illustrate the feed-forward connections.

Convolutional Neural Network (CNN)

CNNs are another type of neural networks commonly used in image classification tasks and time series analysis[107]. They are the most used type of deep learning model for behaviour recognition and classification, as they have been successfully applied to human activity recognition[165–167] and AAR[75,76,90,91,109–113,163] from accelerometer data. CNNs hierarchically apply filters of increasing complexity which helps to automatically capture patterns in the imagery or from raw input data[76,108]. The output of each convolutional layer is a feature map of the input image or of the previous layer's output. Usually, the output of a convolutional layer is down-sampled through pooling layers, which compute summary statistics, and then are passed on to another convolutional, flattening, or dense layer. Different numbers of filters, kernel sizes and strides can be applied to capture local temporal dependencies. The computational cost and processing times depends on the complexity, depth, and width of the architecture[168]. CNNs usually do not need preceding feature extraction, as they work directly on the raw data or images of the accelerometer time series, but there have been reports on superior performance of CNNs using statistical features as input compared to raw sensor data[75,90,110]. The best performing model architecture based on an input layer, followed by three blocks of Conv2D/MaxPooling, each with kernel size = 2. The Conv2D layers increased in width with depth of the model with 32, 64 and 128 and 256 filters respectively. The output of the last Conv2D was flattened and passed to a Dense layer with 200 units, a Dropout layer (rate = 0.25), another Dense layer with 100 units and finally the output layer with 3 units. The activation functions were ReLu for the convolutional blocks and Dense layers, and softmax for the output dense layer. The applied optimizer was the Adaptive Moment Estimation (Adam) with a learning rate schedule and the loss function was the Categorical Cross-Entropy. The model was trained over 50 epochs using a 10-fold cross-validation, so 5 epochs per fold.

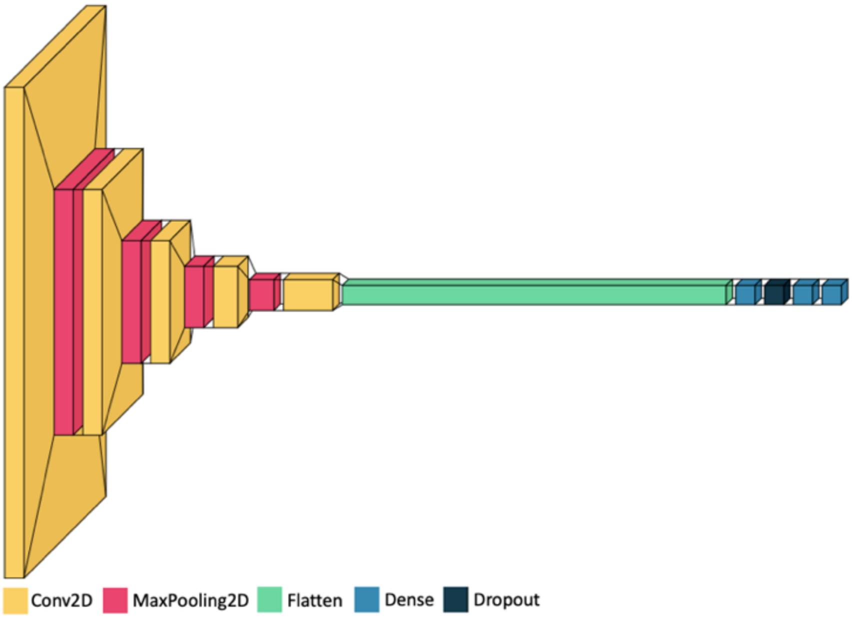


Figure 10: 3D representation of the CNN architecture

Autoencoder

Autoencoders[169] are artificial neural networks usually used for unsupervised image reconstruction. They consist of an encoder and a decoder part. The encoder compresses the input data into a low-dimensional representation containing the most important information in an encoded form. The decoder then reconstructs the original input data from just this encoded representation[170,171].

To create an autoencoder that best possibly extracts the information from the images, an image reconstruction approach was applied. During training, the autoencoder learns to encode the essential features of the input images into a lower-dimensional latent space representation and then decodes it back to reconstruct the original image. The encoder part of the autoencoder learns to capture the most important features of the input images, while the decoder part learns to generate images that closely resemble the original input. By minimizing the reconstruction error between the input and the reconstructed output, the autoencoder learns to extract meaningful features from the input data. For a classification problem, the encoder's saved weights of the best performing autoencoder were combined with additional layers, as fully-connected dense layers, creating a classifier. The more robust representation of the original input images created by the encoder part supposedly is easier to classify[171]. The best performing encoder part consisted of 3 blocks of a

Convolutional layer with 32 filters and a MaxPooling layer respectively. The trained encoder was combined with a BatchNormalization, a Dropout (rate = 0.25), Flatten, Dense (128 neurons), Dropout (rate = 0.25), Dense (64 neurons), Dropout (rate = 0.25), a BatchNormalization and the final Output Dense layer with 3 neurons and a softmax activation function. The model was trained for 70 epochs, with 7 epochs per fold.

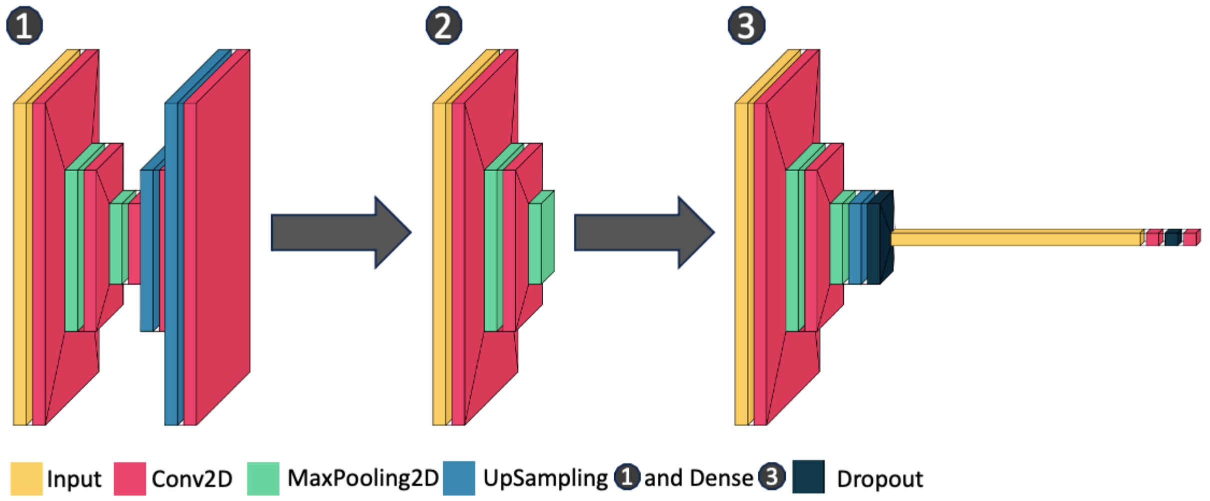


Figure 11: (1) Autoencoder encodes and reconstructs input (2) The trained encoder part (3) Encoder combined with a classifier.

Residual Neural Network (ResNet)

ResNets[172] are a type of deep neural network architectures with a residual learning mechanism allowing the very deep networks with hundreds to thousands of layers without a vanishing gradient. This depth enables capturing complex features and patterns. Their main innovations are the residual learning and shortcut connections. The residual learning is based on using residual blocks instead of traditional stacked layers. The input to the block is added to the output of the block, rather learning a residual mapping instead of directly learning the feature mapping. The shortcut connections skip one or more layers, allowing the gradients to flow directly through the network[173]. Even though ResNets have proven to be among the most accurate image and time-series classification algorithms[174], several applications in AAR reported that ResNets tend to overfit regardless of the choice of hyperparameter values[154,174,175]. Potentially caused because by the depth that allows memorization of uninformative and irrelevant patterns in the training data[3]. The applied optimizer was the Adaptive Moment Estimation (Adam) with a learning rate schedule and the loss function was the Categorical Cross-Entropy. The model was trained for 60 epochs, with 6 epochs per fold.

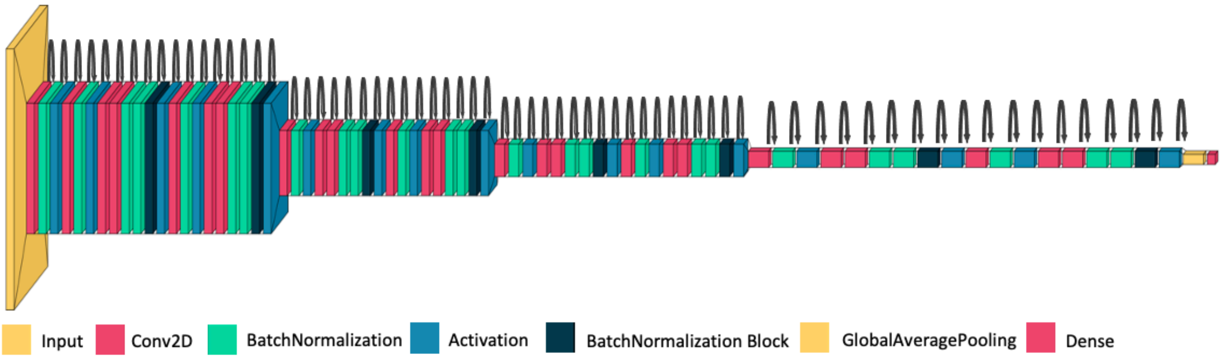


Figure 12: 3D representation of the ResNet with arrows as shortcut skip connections between layers.

Triple Cross-Domain Attention CNN (TCDA-CNN)

The TCDA-CNN is a sophisticated neural network architecture, which has shown great performance in recognizing human activity from accelerometer data. The TCDA-CNN integrates advanced attention mechanisms to capture diverse aspects of the input data[176]. The following methodology was applied according to Tang et al., 2022, where the algorithm was applied to human activity recognition.

The raw sensor data input comes in the shape of (Amplitude, Timesteps, Sensor Axes). The network begins with the creation of three parallel branches, which allow focus on dimension interactions. This segmentation enhances the network's ability to focus on relevant information from multiple domains and selectively emphasizes or suppresses specific regions, features or sensor axes. The first branch takes a rotated version of the original input with shape (Timesteps, Amplitude, Sensor Axes), the second another rotated version with shape (Timesteps, Sensor Axes, Amplitude) and the last branch takes the version with shape (Sensor Axes, Amplitude, Timesteps). Then, for each branch a Z Pooling operation calculates the Max and Average across the last dimension creating a shape of (Amplitude, Timesteps, 2), (Timesteps, Amplitude, 2) and (Sensor Axes, Amplitude, 2) respectively. The following Conv2D layers with 1 filter extract the most relevant features from each domain. After applying this attention mechanism, the three outputs are fused to obtain a unified representation that incorporates information from all domains. This is done by back rotating all three lanes to the original shape (Amplitude, Timesteps, Sensor Axes) and combining them to a feature map by weighting each input by 1/3. This fused representation can then be used for further processing, such as classification. By incorporating the triple cross-domain attention mechanism into a classifier, the network becomes more capable of capturing relevant information from multiple domains or sources. This can lead to improved performance, as the network becomes better at focusing on informative features and suppressing noise or irrelevant information. The best performing architecture was the TCDA mechanism followed by a Flatten layer, then a Dense layer with 128 neurons, a Dropout layer (rate = 0.25), another Dense layer with 64 neurons and finally an output layer with 3 neurons and a softmax activation function. The applied optimizer was the Adaptive Moment Estimation (Adam) with a learning rate schedule and the loss function was the Categorical Cross-Entropy. The model was trained over 70 epochs, with 7 epochs per fold.

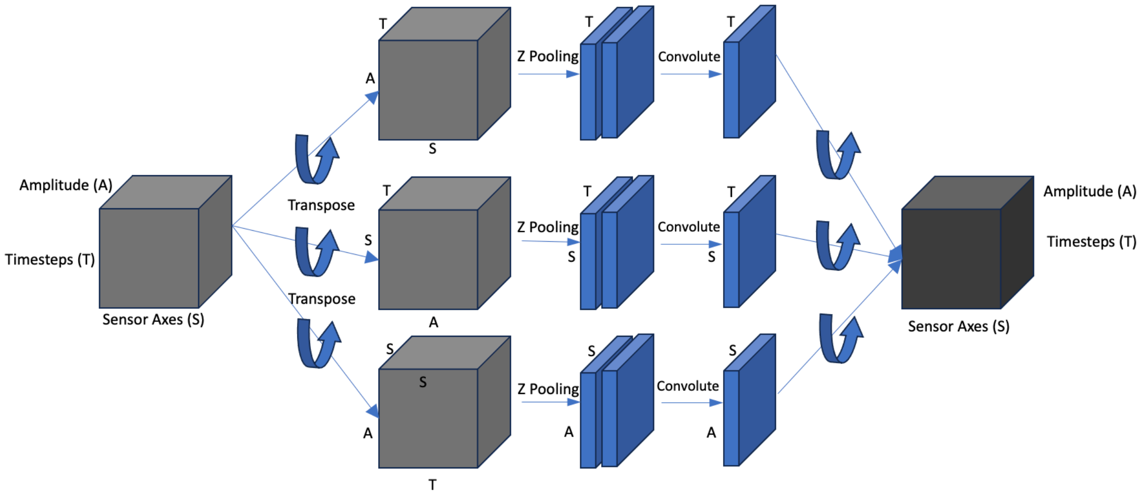


Figure 13: Input is rotated into 3 branches, then Z pooled [Max, Average], Convolution (1 filter), Fusion by weighting each Conv2D output by 1/3.

Recurrent Neural Networks and Long Short-Term Memory (LSTM)

RNNs are neural networks with loops that allow information to persist over time and hence are particularly effective for modelling sequential data, such as time series, natural language, and audio signals. They have found successful application in AAR[95,96,98,162,177]. They process sequences of inputs by iterating through the sequence elements while maintaining a hidden state that captures information about previous inputs. But RNNs struggle to capture long-term dependencies due to the vanishing gradient problem, where gradients become extremely small during backpropagation, making it difficult to learn from distant past information[178]. The LSTM is a type of RNN architecture designed to overcome this vanishing gradient problem. To address this problem, a more sophisticated architecture with memory cells and gating mechanisms is introduced. The LSTM unit includes an internal cell state (memory) which is iteratively updated over time generating a linear pathway running through the entire sequence allowing information to flow. This allows the understanding of context and the capturing of long-term dependencies. LSTM units contain different gating mechanisms (input gate, forget gate, output gate) that regulate the flow of information through the network, allowing it to selectively retain or discard information based on its relevance. The forget gate decides which information to remove from the cell state. It takes input from the current input and the previous hidden state, producing a forget gate vector that scales the previous cell state. The input gate determines which information to store in the cell state. It consists of two components: an input gate that decides which values to update and a tanh layer that creates a vector of new candidate values. The output gate controls which parts of the cell state are exposed as the output. It combines the updated cell state with the current input and previous hidden state to produce the output. These gating mechanisms allow LSTM units to selectively learn and forget information over time, enabling them to capture longer-term dependencies compared to traditional RNNs[179].

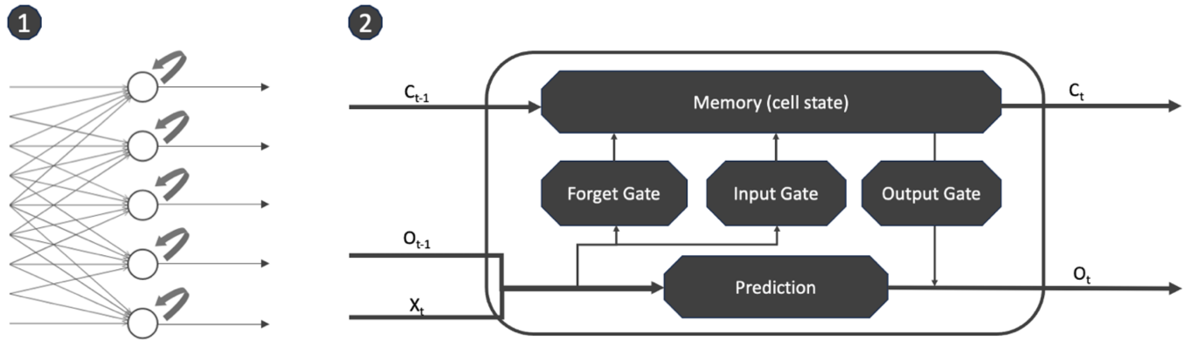


Figure 14: 1. Recurrent Neural Network with memory or feedback (arrows) compared to Feed-Forward Neural Networks. 2. LSTM cell visualization with input from previous layer and output to subsequent layer.

CNN-LSTM

Such LSTMs have been combined with CNNs to form LSTM-CNN or CNN-LSTM hybrid models. The latter hybrid method has been successfully applied to AAR on livestock and pets detecting a broad spectrum of behaviours[3,118,155,156,180–182]. A CNN-LSTM is a hybrid neural network architecture that combines the strengths of CNNs and LSTM networks to classify sequential data, such as time series or sequential sensor data. A CNN-LSTM combination for classification merges the spatial feature extraction capabilities of CNNs and the temporal modelling capabilities of LSTMs to classify sequential data accurately. The input data is pre-processed into fixed-size, overlapping windows and aggregated to batches of consecutive windows. These batches as subsequences of the time series data are fed into the CNN. The CNN layers then extract spatial patterns and local features from the input data. Pooling layers help reduce the spatial dimensions of the features while retaining the most important information. The output of the CNN part is a set of high-level spatial features. The subsequent LSTM layer captures the temporal dependencies and long-range dependencies in the sequential batches. It takes the sequence of spatial features extracted by the CNN as input and processes them sequentially over time, storing and updating information over time. The output of the LSTM is a sequence of hidden states representing the temporal evolution of the input data. The output sequence of hidden states from the LSTM then are fed into additional fully connected dense layers for classification[183].

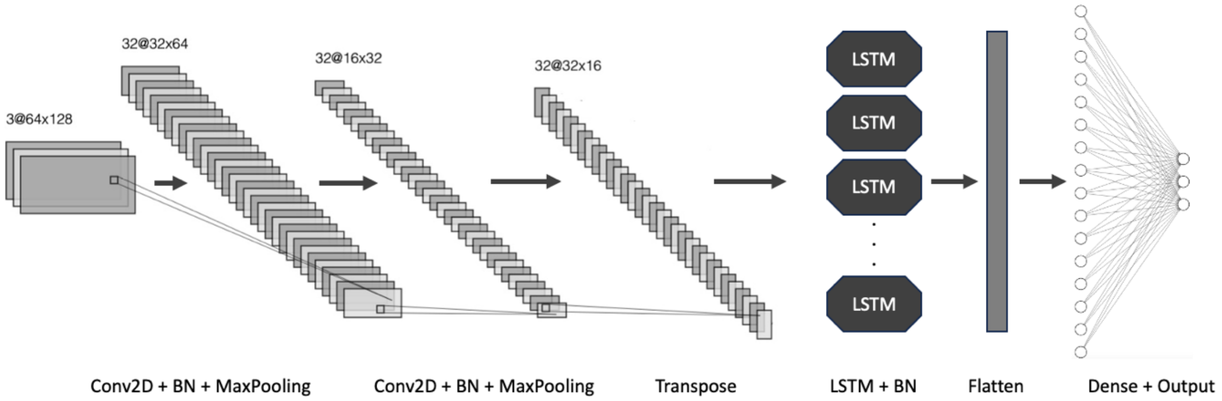


Figure 15: CNN-LSTM architecture visualization with two convolutional, batch normalization and max pooling layers.

LSTM-CNN

A LSTM-CNN is another hybrid model architecture for classification is that integrates LSTM networks with CNNs to classify sequential data. They have performed very well in human activity recognition[184]. The following process is according to Xia et al., 2020. The input data is again pre-processed into fixed-size, overlapping windows and aggregated to batches of consecutive windows. These batches as subsequences of the time series data are fed into the LSTM with n neurons. The LSTM generates a sequence of hidden states with output shape (samples, timesteps, n neurons), which is fed into the CNN. But CNNs can only take four dimensions, so the LSTM output is expanded to shape (samples, 1, timesteps, n neurons). The CNN processes each hidden state independently, extracting spatial features from the temporal representations generated by the LSTM. After processing through the LSTM and CNN layers, the output of the CNN is flattened or pooled to create a feature vector which is then passed through fully connected layers for classification. These fully connected layers aggregate information from the entire input sequence.

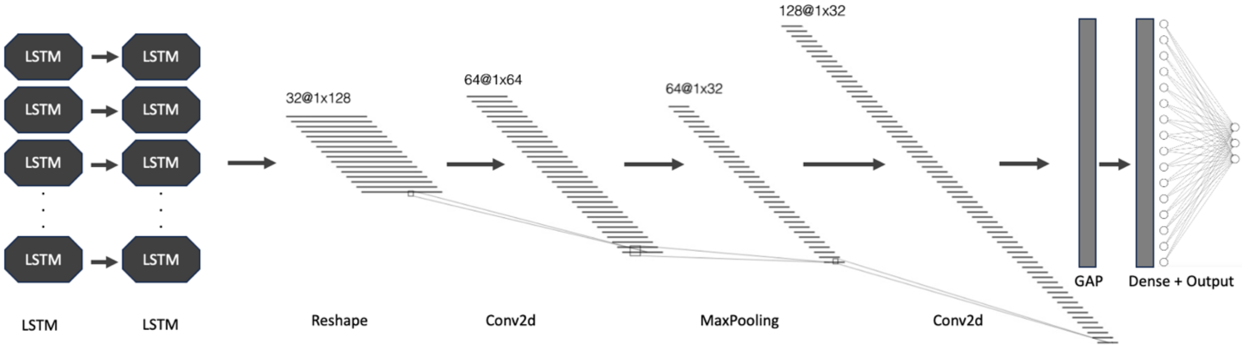


Figure 16: LSTM-CNN architecture with two LSTM layers followed by two convolutional layers, a MaxPooling and a Global Average Pooling layer

Statistical Evaluation

After feeding the labelled raw data or images into different algorithms and training them using a 10-fold cross-validation, adjusting the model architectures and tuning the hyperparameters, the performances were compared. Different combinations of architectures and hyperparameters, including number of layers, filters, optimizers, activation function learning rate, batch normalization, dropout regularization, and skip connections led to small improvements within the same algorithm, without being able to give a consistent pattern. Even though suggested by

Arablouei et al., 2023, using the tanh activation function instead of ReLU, did not lead to noteworthy improvements. Also, more filters and hidden layers did not necessarily lead to better performance, especially on the unseen test set. To circumvent the model learning too well on the training data and thus decreasing generalizability, batch normalization, dropout regularization, and skip connections were considered. To check for overfitting, the training and validation loss was monitored across epochs. Only the very complex (deep and wide) algorithms, started overfitting visually, especially for high numbers of epochs per fold, as they are deep and wide enough to memorize the irrelevant noise in the training data[185]. The advantages of such complexity come with a high nonlinearity and nonconvexity in the optimization functions, making it almost impossible to interpret and analyse the performance of deep learning models[186]. Therefore, really being able to interpret the deep learning models and to understand their performance remain areas of active research[187]. The best performing versions of each algorithm type were compared using visual and numerical metrics. The applied visual metrics were the Receiver Operating Characteristic (ROC) Curve[188], both class-wise and macro-averaged (average of independently computed scores for each class), and confusion matrices[189] for the test set across all cross-validation folds and at the end of training. The used numerical metrics included Matthews correlation coefficient (MCC)[190], the class-wise and macro-averaged Area Under Curve (AUC), recall and precision scores for the test set across all cross-validation folds and at the end of training. The MCC also uses the true and false positives and negatives (TP, FP, TN, FN) and shows values between −1 and +1, where +1 is perfect prediction, 0 no better than random prediction, and −1 perfect inverse prediction. It is known to be a meaningful measure even when the dataset is highly imbalanced[3]. Furthermore, the McNemar's test[191] was used to test the agreement between two classifiers, in this case to compare the baseline MLP classifier with the other classifiers. It was computed using all combined test set predictions across all cross-validation folds and on the test set predictions after training. The metrics were computed using the "statsmodels" (version 0.12.0)[159] and "scikit" (version 0.24.2)[192] libraries. The Welch's t-test[193] was applied to compare the effect of the different window sizes on the performance of the classifiers using the "SciPy" (version 1.73)[194] library.

Analysis of Timeseries & Environmental Variables

The best performing trained and validated algorithm was used to predict unlabelled windows extracted for the years 2022 and 2023 and for individuals with labelled social behaviours. To save computation, time and storage, a selected array of accelerometer readings was downloaded from Movebank to be predicted. For both years, the five days from the 15th to 19th day for each month were selected. For each of these days, 5 minutes for the morning hours from 6:30 AM to 10 AM were downloaded. A shifting mechanism was applied so that for each of these five days for a given month, every new day the sampling period shifted by 10 minutes, so that as much of the hour as possible is covered. The accelerometer readings were named after their timestamps and birdID, which was later used to assign the behavioural predictions to time periods and individuals. Then, several statistics and visualizations were derived from those frequencies, as frequency across the morning hours, development over the years and months. Finally, a dataset of Normalized Difference Vegetation Index (NDVI) and drought timings, provided by Ogina et al., in prep, was included into the analysis, which operates as a proxy for vegetation status and thus food abundance.

Ethics

This study was done under research permits and authorisations from the Max Planck Society *Ethikrat* Committee, the National Commission for Science, Technology and Innovation of Kenya (NACOSTI) and Kenyan Wildlife Service (KWS), as well as in collaboration with the National Museums of Kenya.

Results

Comparison of algorithm performances

Baseline MLP classifier

As expected, the baseline MLP classifier performed better than random choice. The average AUC score and MCC differed between the two window sizes (mean $\text{AUC}_{\text{1second}}$=0.66, mean $\text{MCC}_{\text{1second}}$ =0.29; mean $\text{AUC}_{\text{3seconds}}$=0.62, mean $\text{MCC}_{\text{3seconds}}$ =0.26). The average precision and recall were identical/different between the two window sizes (mean $\text{precision}_{\text{1second}}$ = 0.67, mean $\text{recall}_{\text{1second}}$ =0.44; mean $\text{precision}_{\text{3seconds}}$ = 0.48, mean $\text{recall}_{\text{3seconds}}$ =0.43).

	Window Size	**1 second**				**3 seconds**			
		Classes	BOW	CHA	NOT	**Classes**	BOW	CHA	NOT
Classifier	**Metric**	**Macro Ø**	**Class-wise**	**Class-wise**	**Class-wise**	**Macro Ø**	**Class-wise**	**Class-wise**	**Class-wise**
MLP	AUC	0.668	0.670	0.653	0.623	0.621	0.719	0.456	0.633
	MCC	0.297				0.263			
	Precision	0.669	0.625	0.400	0.920	0.483	0.444	0.000	0.968
	Recall	0.440	0.196	0.071	0.991	0.434	0.286	0.000	0.993
CNN	AUC	0.829	0.834	0.835	0.859	0.819	0.878	0.649	0.797
	MCC	0.396				0.287			
	Precision	0.612	0.655	0.276	0.941	0.460	0.286	0.000	0.972
	Recall	0.515	0.373	0.286	0.963	0.518	0.429	0.000	0.976
TCDA-CNN	AUC	0.785	0.807	0.779	0.815	0.842	0.830	0.718	0.856
	MCC	0.334				0.265			
	Precision	0.537	0.438	0.313	0.932	0.492	0.412	0.083	0.977
	Recall	0.488	0.275	0.179	0.971	0.497	0.500	0.118	0.959
Autoencoder	AUC	0.852	0.835	0.845	0.865	0.862	0.768	0.724	0.772
	MCC	0.450				0.257			
	Precision	0.660	0.600	0.333	0.949	0.466	0.219	0.071	0.977
	Recall	0.505	0.412	0.214	0.963	0.465	0.500	0.118	0.883
ResNet	AUC	0.826	0.817	0.825	0.828	0.851	0.961	0.768	0.872
	MCC	0.357				0.275			
	Precision	0.575	0.395	0.321	0.939	0.445	0.381	0.214	0.979
	Recall	0.523	0.333	0.321	0.949	0.523	0.571	0.176	0.972
CNN-LSTM	AUC	0.857	0.861	0.860	0.846	0.873	0.883	0.743	0.892
	MCC	0.430				0.313			
	Precision	0.614	0.500	0.333	0.954	0.458	0.244	0.162	0.989
	Recall	0.569	0.529	0.357	0.947	0.605	0.714	0.353	0.920
LSTM-CNN	AUC	0.828	0.809	0.803	0.819	0.841	0.932	0.732	0.814
	MCC	0.280				0.276			
	Precision	0.421	0.393	0.500	0.941	0.423	0.318	0.152	0.987
	Recall	0.372	0.431	0.071	0.939	0.485	0.500	0.412	0.880
	Ø AUC	0.806	0.805	0.800	0.808	0.816	0.853	0.684	0.805
	Ø MCC	0.363				0.276			
	Ø Precision	0.584	0.515	0.354	0.940	0.461	0.329	0.098	0.978
	Ø Recall	0.487	0.364	0.214	0.960	0.504	0.500	0.168	0.941

Table 3: Average metrics' scores for each classifier and window size, as suggested in Riaboff et al., 2020 and 2022[89,144]

CNN classifier

The CNN performed much better than the baseline classifier, with high macro-averaged and balanced class-wise AUC scores especially for the 1 second window size. Looking at the confusion matrix (Appendix 3), it is apparent that the CHA class is classified much better than in the MLP (Appendix 1). The McNemar's test only showed some significant disagreements between the baseline MLP and the CNN classifier especially for the 3 seconds window size (Appendix 15).

TCDA-CNN classifier

The TCDA-CNN classifier performed similarly for both window sizes, according to the visual and numerical metrics' scores (Appendix 5 and 6). The McNemar's test results disclosed more significant differences for the 3 seconds window size (Figure 18). The TCDA-CNN showed more confusion between the two social behaviour classes than the CNN (Appendix 5 and 6). Particularly for the 3 seconds window size the CHA class gets misclassified a bit more than for the 1 second window size.

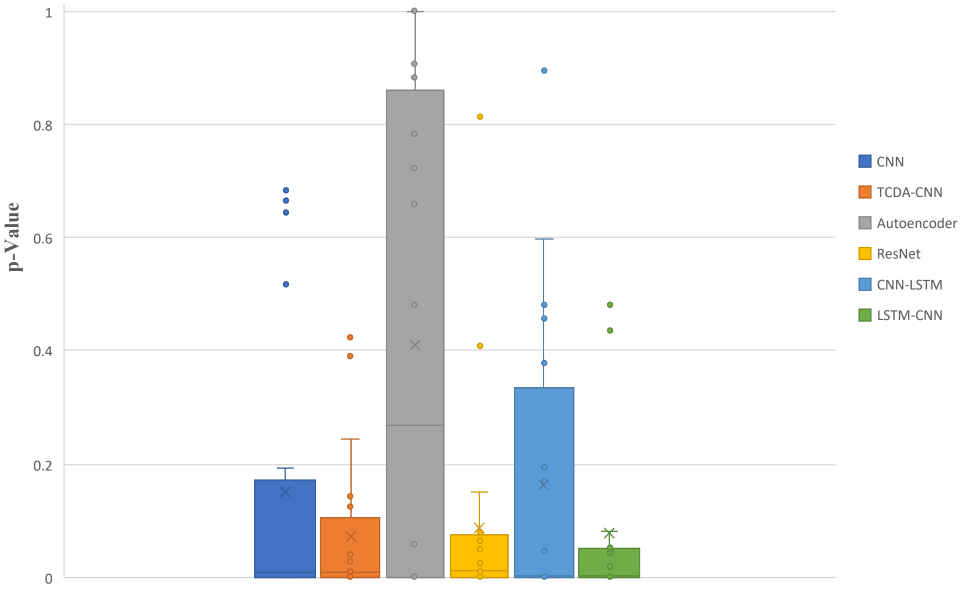


Figure 17: McNemar's test results, where the predictions of each algorithm are compared to the baseline MLP classifier for both window sizes. If the p-Value is >0.05, then, the null hypothesis, that there is no difference in the misclassification patterns between the two models, can be rejected.

Autoencoder classifier

The Autoencoder achieved balanced AUC scores across classes results for both window sizes, being higher for the 1 second window size (Table 3). The visual metrics disclosed that the 3 seconds window size creates more confusion between social behaviour classes than the 1 second window size (Appendix 7 and 8). The McNemar's test showed almost no significant differences in misclassification patterns compared to the baseline classifier for the 1 second window size, but many for the 3 seconds window size, even though most of the metrics were significantly better for the 1 second window size (Table 4).

ResNet classifier

The ResNet scored balanced scores across classes for the 1 second window sizes. The AUC score for the BOW class was very high ($AUC_{3seconds}$= 0.961), but lower for the CHA class ($AUC_{3seconds}$= 0.768). The McNemar's test again revealed more significant differences for the 3 seconds window size across folds, while the performances were not consistently better for one window size (Table 4). The visual metrics showed some confusion between the social behaviour classes (Appendix 9 and 10).

Classifier	Metric	t	p-Value
MLP	AUC	3.0357	0.0071
	MCC	1.3910	0.1832
	Precision	5.4164	0.0002
	Recall	0.5062	0.6190
CNN	AUC	0.6570	0.5260
	MCC	5.8755	0.0000
	Precision	7.5651	0.0000
	Recall	-0.1255	0.9016
TCDA-CNN	AUC	-3.3209	0.0046
	MCC	3.2244	0.0049
	Precision	2.4723	0.0239
	Recall	-0.5738	0.5748
Autoencoder	AUC	-0.6065	0.5566
	MCC	9.9067	0.0000
	Precision	9.6441	0.0000
	Recall	1.9650	0.0678
ResNet	AUC	-2.6706	0.0165
	MCC	2.5229	0.0240
	Precision	3.8067	0.0022
	Recall	-0.0041	0.9968
CNN-LSTM	AUC	-1.1427	0.2710
	MCC	3.5053	0.0027
	Precision	4.9363	0.0003
	Recall	-1.3795	0.1849
LSTM-CNN	AUC	-0.4499	0.6620
	MCC	0.0951	0.9253
	Precision	-0.0394	0.9690
	Recall	-2.1572	0.0475

Table 4: Pairwise Welch's t-test comparing the performance of the classifier with two different window sizes (1 to 3 seconds). The further the t-statistic away from zero in both directions, the greater the difference between the means of the samples (metrics across all CV folds). If t-statistic is positive, the mean of the 1 second sample is higher, if negative the mean of the 3 seconds sample. A small value (<0.05) indicates strong evidence that the null hypothesis, that there is no difference between the means of the two samples.

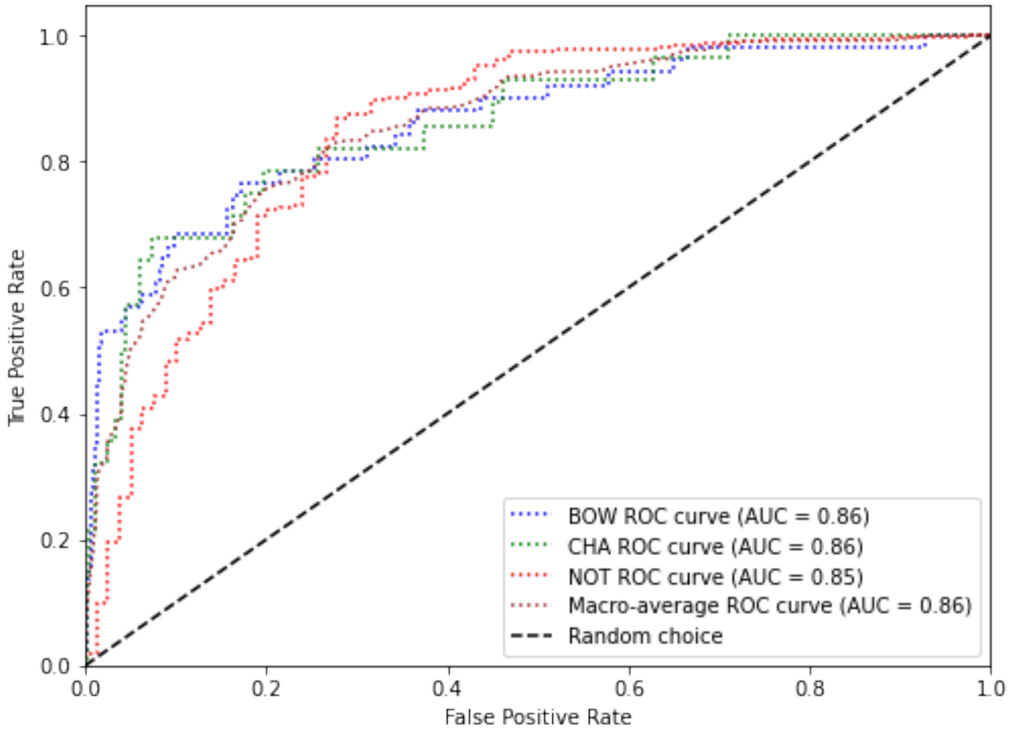


Figure 18: CNN-LSTM ROC curve at the end of training for the 1 second window size

CNN-LSTM classifier

The CNN-LSTM performed very well for both window sizes with balanced metrics scores across classes, being higher for the 1 second window size (Table 4). The McNemar's test disclosed a broad range of p-values for the 1 second window size, while the 3 second window was again significantly different than the MLP classifier, but the two significant differences in the Welch's test promoted the 1 second window size. The visual metrics for the 3 seconds window size manifested confusion even for the NOT class, which was not the case for the 1 second window size at all (Appendix 12 and 13).

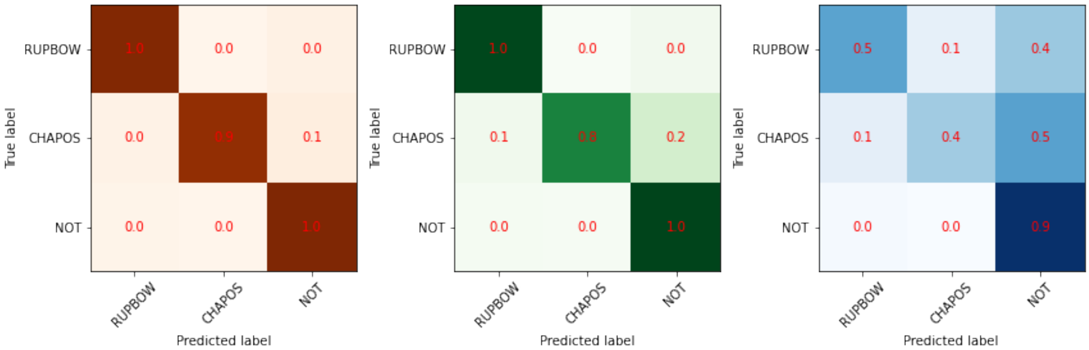


Figure 19: CNN-LSTM confusion matrix at the end of training using the 1 second window size, for the training, validation and test set respectively (left to right). RUPBOW refers to the BOW class, CHAPOS to CHA.

LSTM-CNN classifier

The LSTM-CNN also showed some very good and balanced results for both window sizes (Table 3). The McNemar's test revealed once more significant differences for the 3 seconds window size while the Welch's t-test indicated better and more significant metrics for the 1 second window size (Table 4). The visual metrics showed quite some confusion for both window sizes (Appendix 13 and 14).

Comparison of Performances for Different Window Sizes

There were a few significant differences in the Welch's t-test for the different window sizes within an algorithm class (Table 4). In some cases, one window size performed better on some metrics and vice versa for other cases. There were no consistent patterns to be found, where one window size significantly performed better across all metrics. Only the MLP showed consistently better metrics for the 1 second window size, which were not significant all over. For the other algorithms, there was no consistency found across all metrics. These findings were supported by the McNemar's test comparing the classifiers also for different window sizes (Appendix 15), where the 3 seconds window size for most algorithms seemed to be consistently and significantly different from the baseline MLP classifier. The generalised performance (bottom Table 3) across all classifiers, the mean metrics' scores were similar between the two window sizes but slightly in favour of the 1 second window size.

Ecological Results

The best performing model, the CNN-LSTM, was used to predict the unlabelled data across the two selected years, during five days every month for the hours 6, 7, 8, 9 and 10 AM. There were 48’724 windows available for 2022 and 168’283 for 2023, totalling 217’007 windows across both years (Table 5). There was much less data available for the year 2022 compared to 2023, as well as for the hours 9 and 10 AM, which can be explained by the inconsistencies in the employment of the tags across time. For January 2022 there was no data available at all for the focal birds. In general, the number of samples is varying quite a bit across years but also across months within the same year (Table 7). This inconsistency in the number of samples could potentially introduce a sampling variability, where smaller sampling sizes are not as representative for the respective period or individual. For the individual birds, there was also a big variability in available windows to predict, especially across years, where the differences were enormous at times (Table 8).

	# of Samples
2022	48724
2023	168283
total	217007

Table 5: Number of samples across the two years

	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM
# of Samples 2022	13129	12680	20057	2028	830
# of Samples 2023	42762	41483	53477	24646	5915
total # of Samples	55891	54163	73534	26674	6745

Table 6: Number of windows derived from available data for each hour and year

	# of Samples 2022	# of Samples 2023	total # of Samples
January	0	11353	11353
February	1181	11571	12752
March	1184	7594	8778
April	883	13112	13995
May	5036	24804	29840
June	3818	19989	23807
July	3806	17987	21793
August	3851	9728	13579
September	3821	14799	18620
October	2392	14834	17226
November	11383	9710	21093
December	11369	12802	24171

Table 7: Number of windows derived from available data for each month

	W1430	WT00043	WT00162	W1413	WT00044
# of Samples 2022	0	0	16764	0	31960
# of Samples 2023	33946	50983	61692	14846	6816
total # of Samples	33946	50983	78456	14846	38776

Table 8: Number of available windows for each bird and year

Keeping these limitations in mind, the analysis of the social behaviour frequencies was conducted. The predicted social behaviour events (BOW and CHA) are indicated as the proportion of the total predicted events for a given hour and day. They are henceforth called *frequency* and are given in percentage of total predictions. It was not possible to simply indicate the number of predicted events, as the number of available accelerometer readings varies across time and individuals and thus would not have been of any meaning.

Looking at the social behaviour frequencies for the different individuals across the morning hours and the different seasons, there were some inter-individual differences to be found (Figure 25). The drought reached from June 2022 to March 2023, followed by a wet season until end of May and an intermediate season until mid of October. It should be noted that for the year 2022 there were only two birds available.

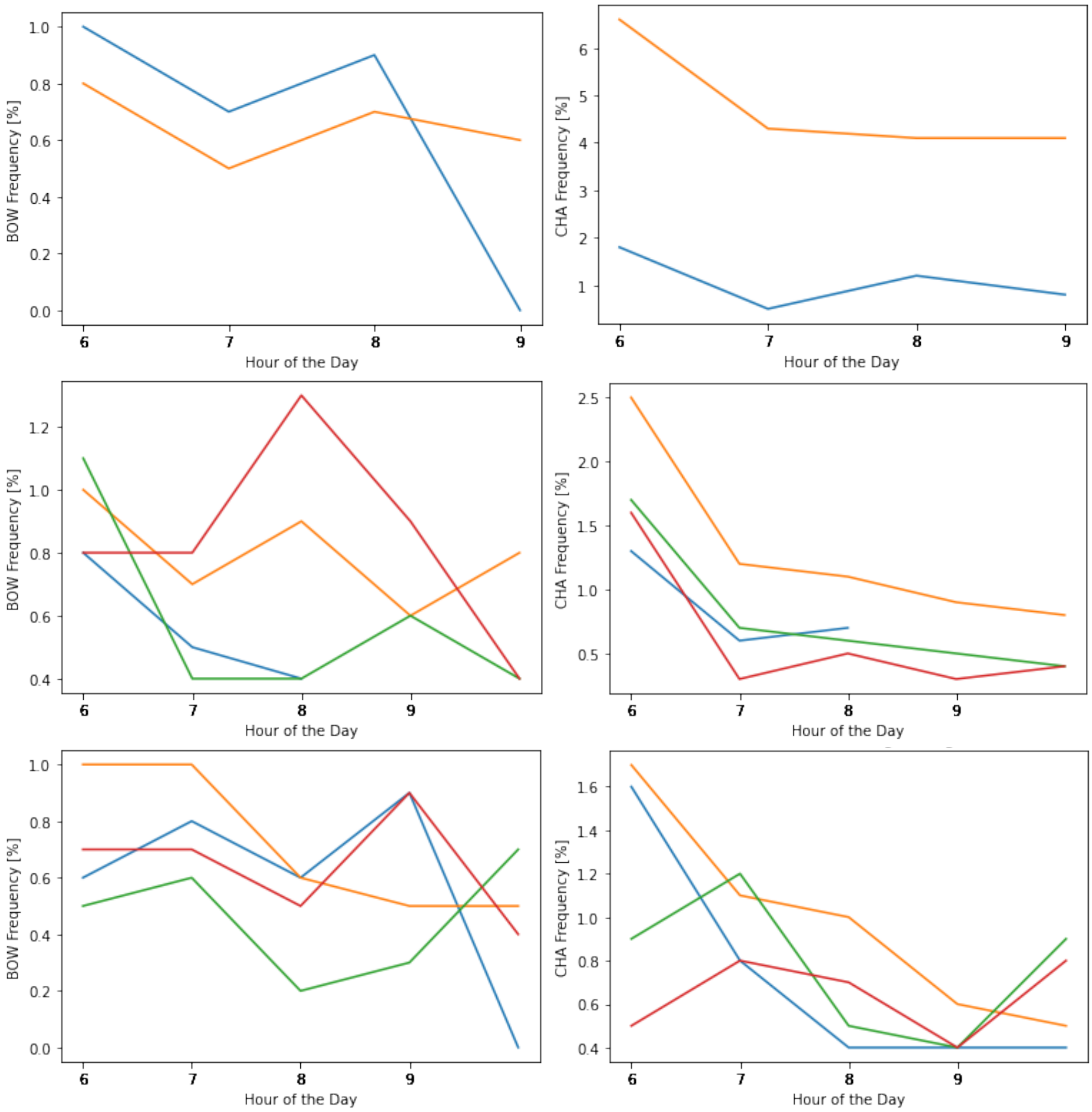


Figure 20: Individual differences in the frequencies during the morning hours for the drought period (top row), the wet seasons (middle row) and the intermediate seasons (bottom row).

In Figure 26 the observed frequencies across the morning hours can be seen. They are plotted together with the increasing mean temperature. The bowing frequencies seem to have a peak around 8 AM, while chasing frequencies decline towards midday. But the scatterplots of temperature with BOW and CHA frequencies resulted in correlation coefficients of 0.04 and -0.16 with p-values of 0.6 and 0.02 respectively.

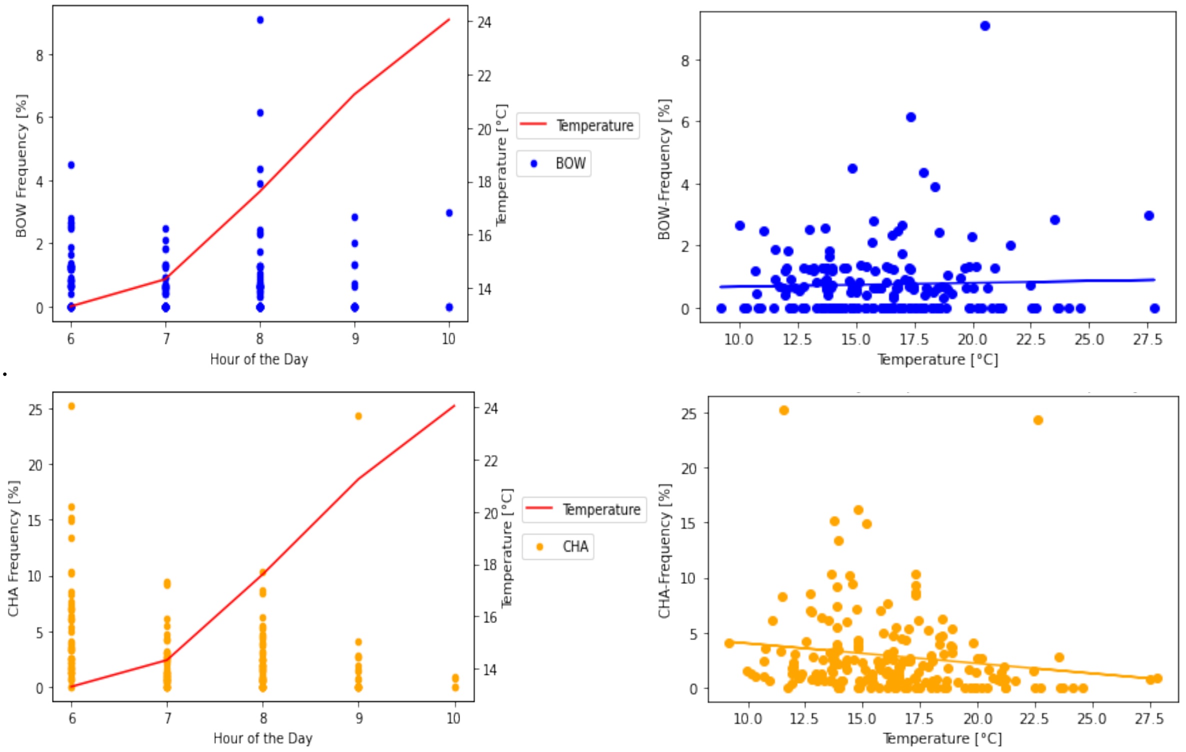


Figure 21: Development in bowing (top left) and chasing (top right) frequencies during the morning hours plotted with the increasing temperature. Scatterplot mean hourly temperature vs. BOW (bottom left) and CHA (bottom right) frequencies, with Pearson correlation coefficients of 0.04 and -0.16 , and p-values of 0.6 and 0.02 respectively.

Plotting the aggregated monthly frequencies across both years, a clear decreasing trend in chasing can be observed, while bowing seems to show no clear trend (Figure 27). The inverse developments of the number of available datapoints and the chasing frequencies across time was obvious (Figure 28). The correlation coefficients for the number of samples and the BOW and CHA frequencies were -0.3 and -0.43 with p-values of 0.16 and 0.04 respectively and indicated small correlations.

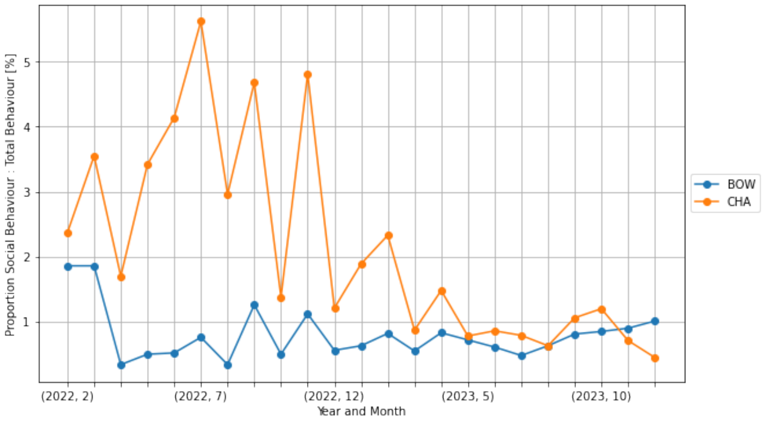


Figure 22: Monthly frequencies for 2022 and 2023

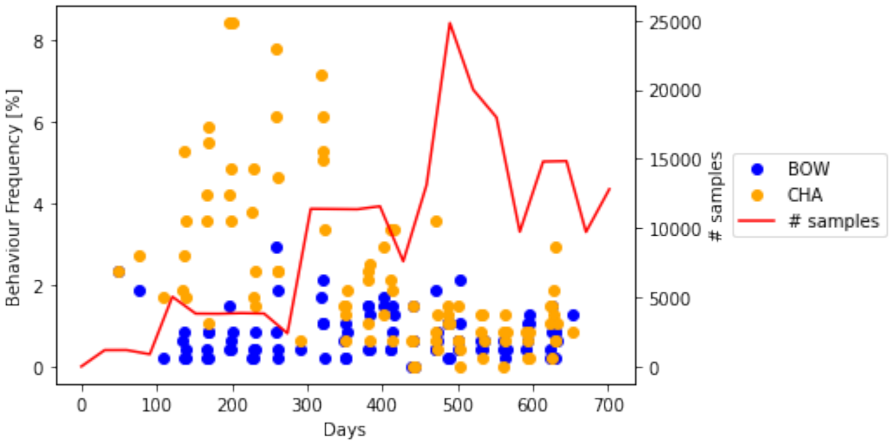


Figure 23: Number of available samples and the predicted social behaviour frequencies

Furthermore, the correlation between BOW and CHA frequencies was analyzed. The Pearson correlation and the scatterplot showed a very weak but significant correlation with a very weak correlation coefficient of 0.15 and a significant p-value of 0.04 (Figure 29).

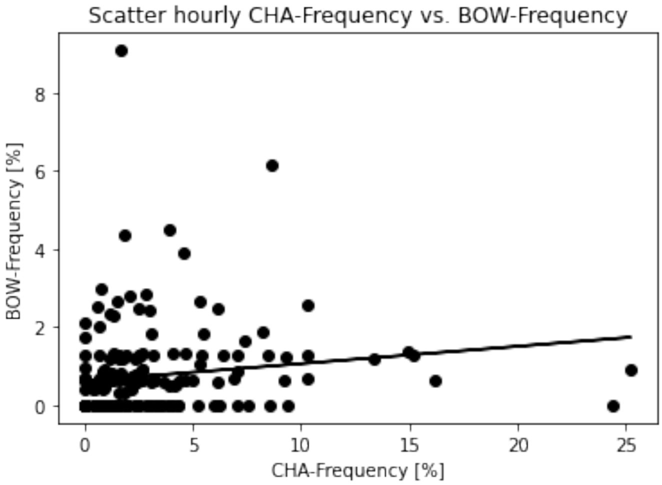


Figure 24: Scatterplot CHA vs. BOW frequencies

To analyze a potential environmental predictor for the occurrence of the social behaviours, the NDVI was consulted. The continuous NDVI timeseries at daily temporal resolution was first plotted together with the predicted frequencies across both years to get a first visualization, but there were no evident correlations (Figure 30).

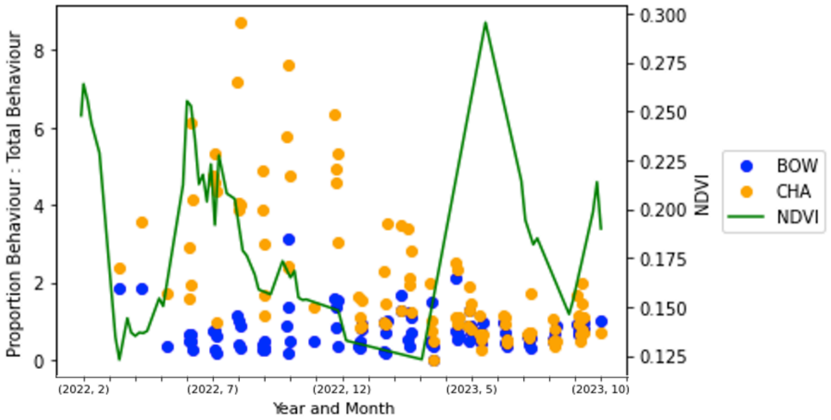
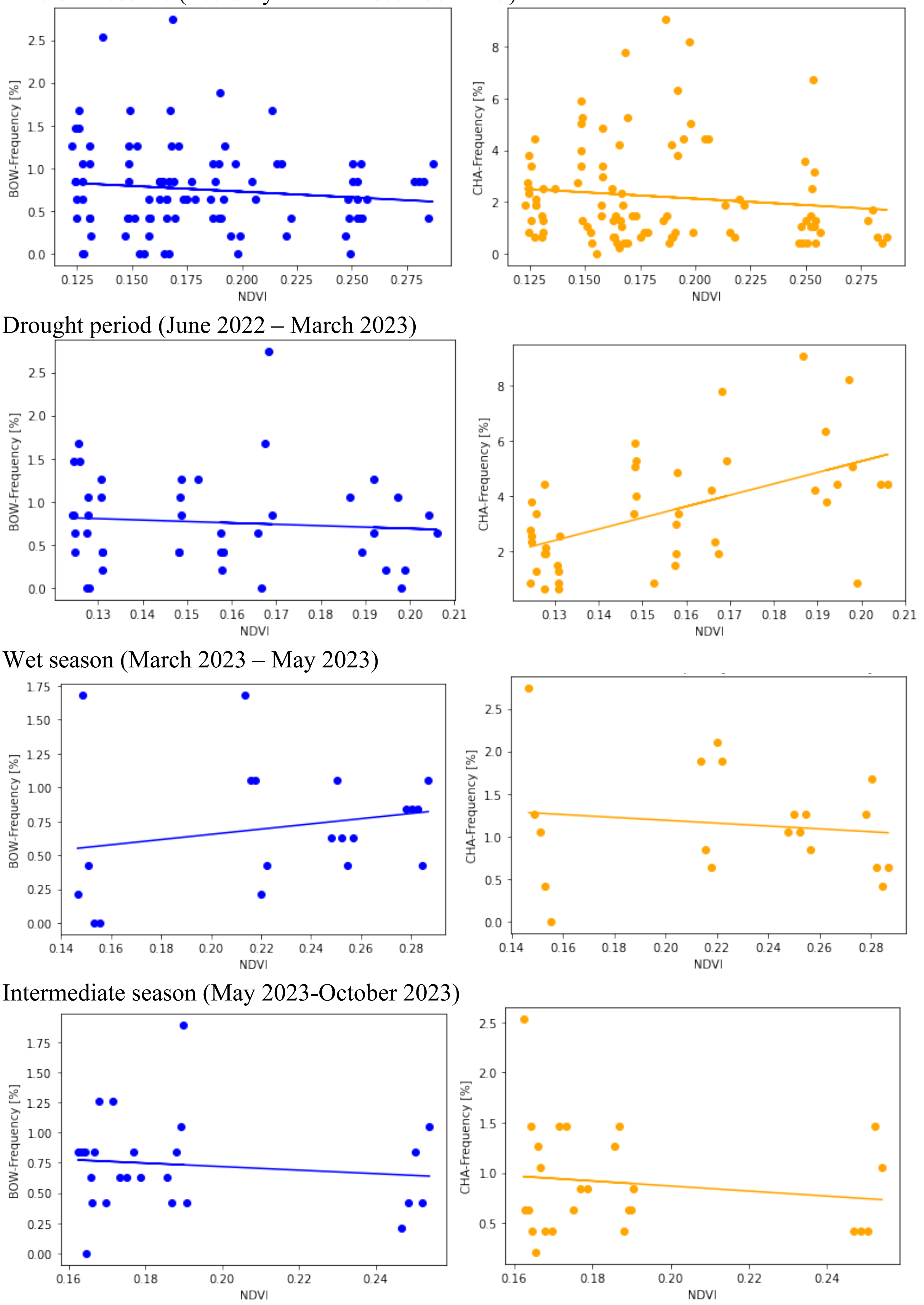


Figure 25: Continuous NDVI timeseries and fragmented daily aggregations of social behaviours

To verify these visual inspections, the daily frequencies of the social behaviours and the daily NDVI underwent a Pearson correlation test across both years and for the three seasons individually. The correlation coefficients between NDVI and BOW and CHA across both years were very weak, with -0.16 and -0.12 and insignificant p-values of 0.1 and 0.22 respectively (Figure 31 top row). The correlation coefficient for the drought period was -0.08 with a p-value of 0.6 for NDVI and BOW. For NDVI and CHA a moderate correlation coefficient of 0.53 with a significant p-value of 0.0002 was found (Figure 31 second row). During the wet seasons, the NDVI and BOW showed an insignificant (p-value = 0.38) 0.2-correlation (Figure 31 third row) and an insignificant (p-value = 0.59) -0.12-correlation for NDVI and CHA. For the intermediate season the NDVI-BOW correlation was again insignificantly (0.55) weakly negative (-0.12), showing a similar pattern as the NDVI-CHA correlation with a coefficient of -0.15 with a p-value of 0.46 (Figure 31 fourth row).


Whole timeseries (February 2022 – December 2023)

Drought period (June 2022 – March 2023)

Wet season (March 2023 – May 2023)

Intermediate season (May 2023-October 2023)

Figure 26: NVDI vs. BOW and CHA across whole timeseries (top row), drought period (June 2022 to March 2023) (2nd row), wet season (March 2023 to end of May 2023) (3rd row) and intermediate season (May 2023 to October 2023) (last row)

Discussion

CNN-LSTM as the Best Performing Classifier

This thesis disclosed the ability of deep learning algorithms to successfully recognize social behaviours in vulturine guineafowl from 20 Hz triaxial accelerometer data. The best performing classifier was the CNN-LSTM architecture using a sliding 1 second window size with 50% overlap. This algorithm's ability to memorize longer-term dependencies was very useful in this sequential data problem. The classifier was chosen due to its low confusion of the social behaviours which show a high inter-activity similarity especially considering the subsequent task of classifying unlabelled data as precisely as possible. Many algorithms performed similarly well, each with their own (dis)advantages and thus present an array of tools to choose from.

Effect of Window Size, Model Complexity and Inter-Activity Similarity

The results suggested that the window sizes did have a minor and varying impact on the model performances. The effect of window sizes depended on the classifier choice, as some performed better with one window size and some with the other. Also, the choice of window size cannot satisfy both class-wise performances equally because the two social behaviours show different durations (Table 1).

Looking at the McNemar's test results, an increase in the complexity of the algorithms did not always bring improvements in performance, especially for the less complex input window size of 1 second, for which basically all algorithms showed only few significant differences to the simple MLP. Another obvious finding is, that the models do not generalize very well on the unseen test set. Many models performed extremely well on the training and validation set, but then tended to misclassify the unseen test set at higher rates. This suggests overfitting models which memorize noise in the data, even though overfitting was not visible in the loss function monitoring over the training process[195]. Looking at the confusion matrices it is apparent that confusion occurs mainly for the test set, so the task at hand is to find models that generalize better. Usually, simpler models tend to generalize better, as they do not learn too much on the training set. The NOT class was never misclassified into the social behaviour classes for the test set, except for the LSTM-CNN classifier (Appendix 13 and 14). This is important to not overestimate the occurrence of social behaviours in the prediction of unlabelled accelerometer data. It makes sense that the two social behaviours do get confused at some rate, as their motion patterns show a high inter-activity similarity[76]. The misclassification of the social behaviour classes into the NOT class might be caused by the inter-activity similarity. This issue is intensified if the dataset does not only include characteristic windows for a given social behaviour class but also vague sequences. Some windows are visually very hard to distinguish from NOT but have been included because the dataset suffers annotation scarcity. Currently, the variation in duration and inter-individual expression is very large for such a small dataset. Thus, splitting into different window sizes thus can strongly diminish the sample size of the characteristic labels. The meticulous work to prepare a valid dataset by cleaning the raw data, handling data gaps and precisely checking the resulting timestamps of the annotation program were crucial steps during pre-processing but further decreased the number of training samples.

Annotation Scarcity and Class Imbalance

This scarcity for the social behaviour classes made the training dataset highly imbalanced. Balancing the datasets with different data augmentation methods, as class-weighting by the inverse frequency and under-sampling the over-represented class were applied here. Nevertheless, the class imbalance is not the main causer of the lower performance of classifying such behaviours, but the scarcity of training data for these behaviour classes itself. A reliable classification of social behaviour is very challenging due to this limited amount of valid training data available[3,155] and establishing accurate classification models with strongly underrepresented behaviours will remain a subject of our ongoing research despite the augmentation techniques[3,155].

Predicting an Imbalanced Sample Size of Unlabelled Accelerometer Data Across Time and Individuals

The model performance and analysis of confusion lead to the decision to apply the best performing and trained algorithm to predict unlabelled accelerometer readings. This real-life scenario enabled a first trial to remotely monitor this population's social behaviours outside of direct field observations. Nonetheless, the derived frequencies and trends over time should be enjoyed with caution because the classification and recognition framework still need more sophisticated and thorough verification through identification tasks. The number of windows available to the prediction part was limited by computational resources and thereby did not allow a consistent prediction of social behaviour across individuals and time. The lower number of samples for some individuals and time periods introduces more sampling variability compared to periods and individuals with a larger sample size[196]. This increases the uncertainty in the explanatory power and decreases the representativeness of the predictions. Hence the temporal and individual patterns of the social behaviour frequencies must be consumed with caution.

Ecological Deductions

The social behaviour frequencies show a decrease from the time they leave their roosting sites in the trees towards midday. At the field site close to the equator, it gets hotter very rapidly after sunrise. Derived from direct observations, it is assumed that the vulturine guineafowl, after spending some time on the open glades, they disperse into the bushes to spend the hottest time of the day there. But according to the weak and insignificant correlations of the temperature and social behaviour frequencies, there must be another explanation. Maybe it is rather the location which influences the occurrence of these displays. In the inaccessible and closed surroundings of the bushes, such display potentially is not as effective anymore, as there is not enough space compared to the glades. Another explanation could be that other individuals cannot see the displays in these dense bushes. Also, during such social interactions, concentration might be lacking and thus increase the predation risk [197].

The predicted frequencies of the courtship behaviour across seasons did not align well with the hypothesis and directly observed behaviour in the field season from April to June 2023. It was expected that with growing vegetation and thus food abundance and nesting possibilities[54,127] the courtship behaviour increases in frequency. This was also observed as the rains arrived and the vegetations started to grow, creating the conditions in which vulturine guineafowls opportunistically breed[54]. But in the model predictions there is no evident increase in the frequencies of bowing. Also, the weak positive correlation of the bowing frequencies with the NDVI during the wet season compared to the weak negative correlations during drought and intermediated season did not suffice to fully confirm the hypothesis and observations. Even when with a focus on the periods of strongly increasing NDVI, no correlating changes in the courtship frequencies were found. This could be caused by the poorly generalizing and frequently confusing CNN-LSTM classifier, unable to distinguish bowing from chasing or other behaviours including running. Another possible explanation could be that the NDVI is not always a reliable real-time proxy for the actual vegetation status especially under drought conditions, as the plant adaptations to different drought of different timescales are not yet fully understood[198199,200]. Before March 2023, there has not been a pronounced wet season since 2020. These dry or even drought periods might have had a longer-term effect on the vegetation and thereby decoupled the NDVI patterns from actual vegetation status found by the vulturine guineafowl. Of course, these limitations concerning the NDVI also apply for the dominance interactions. For these, it was hypothesized that the increase in resources allows more male competition as energy does not have to be saved as parsimoniously as during drier and more scare seasons[69]. But, the NDVI, as a proxy of plant primary productivity directly correlated to resource availability[128,201], does only weakly correlate with the predicted frequencies. This attenuates again the confirmation of the hypothesis that with more accessible resources such aggressive behaviours increase in frequency. The expected significant changes in frequencies in the male courtship and dominance behaviours[64,68,69,202,203] with the beginning of the rainy seasons could not be confirmed.

Outlook

A primary goal should be to improve the classifier performance by fine-tuning optimal hyperparameters and finding the best architecture, adapted to the problem at hand. Another approach to increase the model performance could be the use of statistical features as inputs compared to raw sensor data, even for deep learning. This is especially the case when using hybrid classifiers as CNN-LSTMs or LSTM-CNNs[75,90,110]. The inclusion and combination of more dimensions of information, as environmental or spatial context as well as other sensor modalities as gyroscope or magnetometers, could be tested to better distinguish behaviours by extracting more distinctive features[76,144]. This is not too far-fetched, as some individuals even are equipped with IMUs. But even the best classifier having access to a multi-dimensional data cannot work without a proper training set. For this task with imbalanced classes, the training set should only include unmistakable, characteristic examples of social behaviour labels. The training data should be relieved of confusable labels to not increase the quality of the social behaviour labels. Likewise, the quantity must be increased. Creating new datapoints from the raw data and over-sampling could be options for future studies[76]. An idea for data augmentation, that came up during this study, was to combine social behaviour windows from different overlap percentages, for the same window size of course. In this way, the class imbalance could be decreased while simultaneously increasing the model's robustness, as it learns to identify the social behaviours from different angles.

Another vital approach to enhance the training set is to equip and video-record as many individuals as possible to include more variability[89,144]. This can be achieved by including video-recording into the standard procedure for the morning surveys conducted by the field team. At least every now and then, more video-material should be collected so that more variability over time and individuals can be included. This requires a meticulous planning of the employment times of the accelerometer tags across individuals and time, to efficiently increase the probability of capturing social behaviours on camera. The increase and extension of annotated data would, at a later stage, allow an individual-based and time-stratified splitting into training, validation, and test set. These splitting methods should be considered, to better investigate the generalization ability across time and individuals, the inter-individual variability of the motion patterns and the shifts in sensor orientation[76,89,144]. Furthermore, other overlap percentages as 0%, 25% or 75% in contrast to the applied 50 % could be tested and compared to find the optimal classification approach. The same applies for window sizes, where the range between 1 second and 4 seconds should be tested more precisely (Table 1). Also, considering splitting the whole behaviour sequences into its basic units e.g. pecking, running up, actual bowing and pecking again (Figure 4 and 5). Testing the impacts of increased or decreased sampling rates on the classification performance, would be very interesting[3]. At some point, the effect of including more behavioural classes into the classification problem should be analysed, as for now the NOT class includes many different (state) behaviours outside of these two social behaviour classes. There is yet another identification task to investigate the model's capability of correctly predicting these social behaviours which are exclusive to males. Predicting unlabelled windows for female birds, could check the classifier's plausibility as for these behaviours there should be no predictions. If the females received some positive predictions, this would verify the claim that there are some significant inter-activity similarities. The within-male variability in the male social behaviour frequencies should be compared to the obtained frequency predictions in females. All these potential identification tasks could help to improve and more accurately apply the trained models to real-life scenarios. This includes predicting behaviours for individuals wearing an accelerometer, but not (yet) video-recorded, or predicting behaviours for time periods, not (yet) analysed[144].

But establishing a thorough remote monitoring of the vulturine guineafowl's social behaviour, not only requires a well-performing classifier. The prediction of unlabelled data needs an up-scaled sample size of unlabelled data balanced across time and individuals. In this study the inclusion of more data was limited by the computational and storage possibilities. This would allow a more robust and representative analysis of these behaviours across time and individuals. These first findings call for more investigation e.g. looking into the predictors that influence these behaviours or the patterns underlying the individual expressions of when and how they perform these social behaviours. The causes of the temporal trends in these individually expressed social behaviours could be investigated by including other potential environmental explaining variables such as precipitation and temperature into the model.

Conclusion

To conclude, these findings demonstrated the ability of various deep learning models to classify short and rare social behaviours in a wild, free-ranging population of vulturine guineafowl. Although the algorithms seem to recognize the focal courtship and dominance behaviours quite well, the predicted frequencies from unlabelled data and the deducted ecological analyses should be considered carefully. This methodology still needs some more clarifications to reliably make real-life predictions. However, this methodology with its demonstrated pitfalls could lay the foundation for a long-term remote monitoring of vulturine guineafowl social behaviour. This could be very useful to capture and study the impacts of climate change on behavioural adaptations. But more direct observational data must be collected to establish a high-quality training dataset, that includes more inter-individual and temporal variation. Furthermore, more experimentation with different architectures and hyperparameters distinguishing more behavioural classes to predict should be aimed for.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Damien Farine, for his invaluable guidance, unwavering support, and insightful feedback throughout the duration of this research project. His expertise, patience, and understanding for my situation have been crucial in helping me navigate through the complexities of this work. I am also deeply grateful to my co-supervisors, Dr. Charlotte Christensen, and Dr. André Ferreira, for their mentorship, encouragement, and valuable insights. Their dedication and commitment to academic excellence have inspired me to strive for higher standards in my work. I extend my heartfelt appreciation to Prof. Dr. Robert Weibel, for his practical inputs and hands-on support during the completion of my thesis. His sympathetic guidance and positive feedback significantly contributed to the final realization of this research work. I also want to thank all the members of the Vulturine Guineafowl Project field team, Brendah Nyaguthii, Kennedy Kipkorir, Wismer Cherono, Monicah Wambui and John Wanjala for their collaboration, camaraderie, and intellectual contributions. Their diverse perspectives enthusiasm during discussions have enriched my research experience at Mpala Research Centre. Furthermore, I would like to thank my beloved ones for their unwavering love, encouragement, and support throughout this journey. Their patience, understanding, and belief in my abilities have been a constant source of strength and motivation.

Bibliography

1. Lorenz, K. *The foundations of ethology*. (Springer verlag, 1981).
2. Crook, J. H. & Goss-Custard, J. D. Social ethology. *Annu. Rev. Psychol.* **23**, 277–312 (1972).
3. Arablouei, R. *et al.* Animal behavior classification via deep learning on embedded systems. *Comput. Electron. Agric.* **207**, (2023).
4. Seeley, T. D. & Sherman, P. W. Animal behaviour. *Encyclopedia Britannica* (2023).
5. Beauchamp, T. L. & Frey, R. G. *The Oxford handbook of animal ethics*. (Oxford University Press, USA, 2011).
6. Palmer, C. *Animal ethics in context*. (Columbia University Press, 2010).
7. Rollin, B. E. The regulation of animal research and the emergence of animal ethics: a conceptual history. *Theor. Med. Bioeth.* **27**, 285–304 (2006).
8. Burkhardt, R. W. *Patterns of behavior: Konrad Lorenz, Niko Tinbergen, and the founding of ethology*. (University of Chicago Press, 2005).
9. Tinbergen, N. *The study of instinct*. (Pygmalion Press, an imprint of Plunkett Lake Press, 2020).
10. Buchholz, R. Behavioural biology: an effective and relevant conservation tool. *Trends Ecol. Evol.* **22**, 401–407 (2007).
11. Curio, E. Conservation needs ethologv. *Trends Ecol. Evol.* **11**, 260–263 (1996).
12. Ward, A. & Webster, M. *Sociality: the behaviour of group-living animals*. vol. 407 (Springer, 2016).
13. Rubenstein, D. I. & Rubenstein, D. R. Social Behavior. in (ed. Levin, S. A. B. T.-E. of B. (Second E.) 571–579 (Academic Press, 2013). doi:https://doi.org/10.1016/B978-0-12-384719-5.00126-X.
14. Wey, T., Blumstein, D. T., Shen, W. & Jordán, F. Social network analysis of animal behaviour: a promising tool for the study of sociality. *Anim. Behav.* **75**, 333–344 (2008).
15. Cantor, M. *et al.* The importance of individual-to-society feedbacks in animal ecology and evolution. *J. Anim. Ecol.* **90**, 27–44 (2021).
16. Bralten, J. *et al.* Genetic underpinnings of sociability in the general population. *Neuropsychopharmacology* **46**, 1627–1634 (2021).
17. Godoy, I., Korsten, P. & Perry, S. E. Genetic, maternal, and environmental influences on sociality in a pedigreed primate population. *Heredity (Edinb).* **129**, 203–214 (2022).
18. Faure, J. M. & Jones, R. B. Genetic influences on resource use, fear and sociality. in *Welfare of the laying hen. Papers from the 27th Poultry Science Symposium of the World's Poultry Science Association (UK Branch), Bristol, UK, July 2003* 99–108 doi:10.1079/9780851998138.0099.
19. Crozier, R. H. Genetics of sociality. *Soc. insects* **1**, 223–286 (1979).
20. Wilczynski, W. & Ryan, M. J. The behavioral neuroscience of anuran social signal processing. *Curr. Opin. Neurobiol.* **20**, 754–763 (2010).
21. Parmigiani, S., Palanza, P., Rodgers, J. & Ferrari, P. F. Selection, evolution of behavior and animal models in behavioral neuroscience. *Neurosci. Biobehav. Rev.* **23**, 957–970 (1999).
22. Kumsta, R., Hummel, E., Chen, F. S. & Heinrichs, M. Epigenetic regulation of the oxytocin receptor gene: implications for behavioral neuroscience. *Front. Neurosci.* **7**, 83 (2013).
23. Rubenstein, D. R. Stress hormones and sociality: integrating social and environmental stressors. *Proc. R. Soc. B Biol. Sci.* **274**, 967–975 (2007).
24. Trumbo, S. T. Juvenile hormone and parental care in subsocial insects: implications for the role of juvenile hormone in the evolution of sociality. *Curr. Opin. insect Sci.* **28**,

13–18 (2018).

25. Adkins-Regan, E. *Hormones and animal social behavior*. (Princeton University Press, 2013).
26. Hess, E. H. Imprinting: an effect of early experience, imprinting determines later social behavior in animals. *Science (80-.).* **130**, 133–141 (1959).
27. Robinson, G. E., Fernald, R. D. & Clayton, D. F. Genes and social behavior. *Science (80-.).* **322**, 896–900 (2008).
28. Melzack, R. & Thompson, W. R. Effects of early experience on social behaviour. *Can. J. Psychol. Can. Psychol.* **10**, 82 (1956).
29. Ross, K. G. Molecular ecology of social behaviour: analyses of breeding systems and genetic structure. *Mol. Ecol.* **10**, 265–284 (2001).
30. Seebacher, F. & Krause, J. Physiological mechanisms underlying animal social behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 372 20160231 (2017).
31. Chakravarty, P., Cozzi, G., Ozgul, A. & Aminian, K. A novel biomechanical approach for animal behaviour recognition using accelerometers. *Methods Ecol. Evol.* **10**, 802–814 (2019).
32. Sahin, Y. G. Animals as Mobile Biological Sensors for Forest Fire Detection. *Sensors* vol. 7 3084–3099 (2007).
33. Petersen, J. K. Understanding Surveillance Technologies: Spy Devices, Their Origins \& Applications. in (2000).
34. Banzi, J. F. A Sensor Based Anti-Poaching System in Tanzania National Parks. in (2014).
35. Tibbetts, E. A., Pardo-Sanchez, J. & Weise, C. The establishment and maintenance of dominance hierarchies. *Philos. Trans. R. Soc. B* **377**, 20200450 (2022).
36. Takahashi, M., Tobey, J. R., Pisacane, C. B. & Andrus, C. H. Evaluating the utility of an accelerometer and urinary hormone analysis as indicators of estrus in a zoo-housed koala (Phascolarctos cinereus). *Zoo Biol.* **28**, 59–68 (2009).
37. Fernö, A. Aggressive behaviour between territorial cichlids (Astatotilapia burtoni) in relation to rank and territorial stability. *Behaviour* **103**, 241–258 (1987).
38. Ueda, A. & Kidokoro, Y. Aggressive behaviours of female Drosophila melanogaster are influenced by their social experience and food resources. *Physiol. Entomol.* **27**, 21–28 (2002).
39. Bouissou, M.-F. Androgens, aggressive behaviour and social relationships in higher mammals. *Horm. Res. Paediatr.* **18**, 43–61 (1983).
40. Sandnabba, N. K. Territorial behaviour and social organization as a function of the level of aggressiveness in male mice. *Ethology* **103**, 566–577 (1997).
41. Parmigiani, S. & Pasquali, A. Aggressive responses of isolated mice towards 'opponents' of differing social status. *Ital. J. Zool.* **46**, 41–50 (1979).
42. Ricci, L., Summers, C. H., Larson, E. T., O'Malley, D. & Melloni Jr, R. H. Development of aggressive phenotypes in zebrafish: interactions of age, experience and social status. *Anim. Behav.* **86**, 245–252 (2013).
43. Bartholomew, G. A. & Collias, N. E. The role of vocalization in the social behaviour of the northern elephant seal. *Anim. Behav.* **10**, 7–14 (1962).
44. Crook, J. H. The evolution of social organisation and visual communication in the weaver birds (Ploceinae). *Behav. Suppl.* (1964).
45. da Silva, M. C., Canário, A. V. M., Hubbard, P. C. & Gonçalves, D. M. F. Physiology, endocrinology and chemical communication in aggressive behaviour of fishes. *J. Fish Biol.* **98**, 1217–1233 (2021).
46. Johansson, B. G. & Jones, T. M. The role of chemical communication in mate choice. *Biol. Rev.* **82**, 265–289 (2007).

47. Hurst, J. L. *et al.* Information in scent signals of competitive social status: the interface between behaviour and chemistry. *Chem. Signals Vertebr. 9* 43–52 (2001).
48. Nyaguthii, B. *et al.* Cooperative breeding in a plural breeder: the vulturine guineafowl (Acryllium vulturinum). *bioRxiv* 2011–2022 (2022).
49. Arnold, K. E. & Owens, I. P. F. Cooperative breeding in birds: the role of ecology. *Behav. Ecol.* **10**, 465–471 (1999).
50. Arnold, K. E. & Owens, I. P. F. Cooperative breeding in birds: a comparative test of the life history hypothesis. *Proc. R. Soc. London. Ser. B Biol. Sci.* **265**, 739–745 (1998).
51. Lack, D. Courtship feeding in birds. *Auk* **57**, 169–178 (1940).
52. Ota, N., Gahr, M. & Soma, M. Tap dancing birds: the multimodal mutual courtship display of males and females in a socially monogamous songbird. *Sci. Rep.* **5**, 16614 (2015).
53. Doerr, N. R. Male Great Bowerbirds perform courtship display using a novel structure that rivals cannot destroy. *Emu-Austral Ornithol.* **118**, 313–322 (2018).
54. Nyaguthii, B. *et al.* Cooperative breeding in a plural breeder: the vulturine guineafowl (Acryllium vulturinum). *bioRxiv* (2023).
55. Emery, N. J., Seed, A. M., Von Bayern, A. M. P. & Clayton, N. S. Cognitive adaptations of social bonding in birds. *Philos. Trans. R. Soc. B Biol. Sci.* **362**, 489–505 (2007).
56. Slagsvold, T. & Wiebe, K. L. Social learning in birds and its role in shaping a foraging niche. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 969–977 (2011).
57. Lefebvre, L. & Bouchard, J. Social learning about food in birds. *Biol. Tradit. Model. evidence. Cambridge Univ. Press. Cambridge, United Kingdom* 94–126 (2003).
58. Aikens, E. O., Bontekoe, I. D., Blumenstiel, L., Schlicksupp, A. & Flack, A. Viewing animal migration through a social lens. *Trends Ecol. Evol.* (2022).
59. Candolin, U., Fletcher, R. J. & Stephens, A. E. A. Animal behaviour in a changing world. *Trends Ecol. Evol.* **38**, 313–315 (2023).
60. Wilson, M. W. *et al.* Ecological impacts of human-induced animal behaviour change. *Ecol. Lett.* **23**, 1522–1536 (2020).
61. Tuomainen, U. & Candolin, U. Behavioural responses to human-induced environmental change. *Biol. Rev.* **86**, 640–657 (2011).
62. Buchholz, R. *et al.* Behavioural research priorities for the study of animal response to climate change. *Anim. Behav.* **150**, 127–137 (2019).
63. Conrad, T., Stöcker, C. & Ayasse, M. The effect of temperature on male mating signals and female choice in the red mason bee, Osmia bicornis (L.). *Ecol. Evol.* **7**, 8966–8975 (2017).
64. Gudka, M., Santos, C. D., Dolman, P. M., Abad-Gómez, J. & Silva, J. P. Feeling the heat: Elevated temperature affects male display activity of a lekking grassland bird. *PLoS One* **14**, 1–15 (2019).
65. Jiao, X., Wu, J., Chen, Z., Chen, J. & Liu, F. Effects of temperature on courtship and copulatory behaviours of a wolf spider Pardosa astrigera (Araneae: Lycosidae). *J. Therm. Biol.* **34**, 348–352 (2009).
66. Katsuki, M. & Miyatake, T. Effects of temperature on mating duration, sperm transfer and remating frequency in Callosobruchus chinensis. *J. Insect Physiol.* **55**, 113–116 (2009).
67. Kvarnemo, C. Temperature modulates competitive behaviour: Why sand goby males fight more in warmer water. *Ethol. Ecol. Evol.* **10**, 105–114 (1998).
68. Nguyen, K. & Stahlschmidt, Z. R. When to fight? Disentangling temperature and circadian effects on aggression and agonistic contests. *Anim. Behav.* **148**, 1–8 (2019).
69. Shen, S.-F. *et al.* Unfavourable environment limits social conflict in Yuhina

brunneiceps. *Nat. Commun.* **3**, 885 (2012).

70. Fattorini, N. *et al.* Animal conflicts escalate in a warmer world. *Sci. Total Environ.* **871**, 161789 (2023).
71. Culina, A. *et al.* Connecting the data landscape of long-term ecological studies: The SPI-Birds data hub. *J. Anim. Ecol.* **90**, 2147–2160 (2021).
72. Skinner, B. F. The Evolution of Behaviour 1. in *Behaviour analysis and contemporary psychology* 33–40 (Routledge, 2022).
73. Wynne, C. D. L. & Udell, M. A. R. *Animal cognition: Evolution, behavior and cognition*. (Bloomsbury Publishing, 2020).
74. Bolhuis, J. J., Giraldeau, L.-A. & Hogan, J. A. *The behavior of animals: mechanisms, function, and evolution*. (John Wiley & Sons, 2021).
75. Eerdekens, A. *et al.* A framework for energy-efficient equine activity recognition with leg accelerometers. *Comput. Electron. Agric.* **183**, 106020 (2021).
76. Mao, A., Huang, E., Wang, X. & Liu, K. Deep learning-based animal activity recognition with wearable sensors: Overview, challenges, and future directions. *Comput. Electron. Agric.* **211**, 108043 (2023).
77. Tuyttens, F. A. M. *et al.* Observer bias in animal behaviour research: can we believe what we score, if we score what we believe? *Anim. Behav.* **90**, 273–280 (2014).
78. Brown, D. D., Kays, R., Wikelski, M., Wilson, R. & Klimley, A. P. Observing the unwatchable through acceleration logging of animal behavior. *Anim. Biotelemetry* **1**, 1–16 (2013).
79. Chakravarty, P. *et al.* Seek and learn: Automated identification of microevents in animal behaviour using envelopes of acceleration data and machine learning. *Methods Ecol. Evol.* **11**, 1639–1651 (2020).
80. Setoguchi, S., Kudo, A., Takanashi, T., Ishikawa, Y. & Matsuo, T. Social context-dependent modification of courtship behaviour in Drosophila prolongata. *Proc. R. Soc. B Biol. Sci.* **282**, (2015).
81. Robert, K., Garant, D., Vander Wal, E. & Pelletier, F. Context-dependent social behaviour: Testing the interplay between season and kinship with raccoons. *J. Zool.* **290**, 199–207 (2013).
82. Dougherty, L. R. Meta-analysis shows the evidence for context-dependent mating behaviour is inconsistent or weak across animals. *Ecol. Lett.* **24**, 862–875 (2021).
83. Dawkins, M. S. *Observing animal behaviour: design and analysis of quantitative data*. (Oxford University Press, 2007).
84. Fan, B., Bryant, R. & Greer, A. Behavioral Fingerprinting: Acceleration Sensors for Identifying Changes in Livestock Health. *J* vol. 5 435–454 (2022).
85. Jeantet, L., Vigon, V., Geiger, S. & Chevallier, D. Fully Convolutional Neural Network: A solution to infer animal behaviours from multi-sensor data. *Ecol. Modell.* **450**, (2021).
86. Beliveau, A., Spencer, G. T., Thomas, K. A. & Roberson, S. L. Evaluation of MEMS capacitive accelerometers. *IEEE Des. Test Comput.* **16**, 48–56 (1999).
87. Nathan, R. *et al.* Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures. *J. Exp. Biol.* **215**, 986–996 (2012).
88. Wilson, R. P. *et al.* Give the machine a hand: A Boolean time-based decision-tree template for rapidly finding animal behaviours in multisensor data. *Methods Ecol. Evol.* **9**, 2206–2215 (2018).
89. Riaboff, L. *et al.* Predicting livestock behaviour using accelerometers: A systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data. *Comput. Electron. Agric.* **192**, 106610 (2022).
90. Eerdekens, A., Callaert, A., Deruyck, M., Martens, L. & Joseph, W. Dog's Behaviour

Classification Based on Wearable Sensor Accelerometer Data. in *2022 5th Conference on Cloud and Internet of Things (CIoT)* 226–231 (2022). doi:10.1109/CIoT53061.2022.9766553.

91. Bloch, V., Frondelius, L., Arcidiacono, C., Mancino, M. & Pastell, M. Development and Analysis of a CNN- and Transfer-Learning-Based Classification Model for Automated Dairy Cow Feeding Behavior Recognition from Accelerometer Data. *Sensors* vol. 23 (2023).
92. Walton, E. *et al.* Evaluation of sampling frequency, window size and sensor position for classification of sheep behaviour. *R. Soc. Open Sci.* **5**, 171442 (2018).
93. Dominguez-Morales, J. P. *et al.* Wildlife Monitoring on the Edge: A Performance Evaluation of Embedded Neural Networks on Microcontrollers for Animal Behavior Classification. *Sensors* vol. 21 (2021).
94. Hosseininoorbin, S. *et al.* Deep learning-based cattle behaviour classification using joint time-frequency data representation. *Comput. Electron. Agric.* **187**, 106241 (2021).
95. Peng, Y. *et al.* Dam behavior patterns in Japanese black beef cattle prior to calving: Automated detection using LSTM-RNN. *Comput. Electron. Agric.* **169**, 105178 (2020).
96. Peng, Y. *et al.* Classification of multiple cattle behavior patterns using a recurrent neural network with long short-term memory and inertial measurement units. *Comput. Electron. Agric.* **157**, 247–253 (2019).
97. Shahbazi, M., Mohammadi, K., Derakhshani, S. M. & Groot Koerkamp, P. W. G. Deep Learning for Laying Hen Activity Recognition Using Wearable Sensors. *Agriculture* vol. 13 (2023).
98. Wu, Y. *et al.* Recognising Cattle Behaviour with Deep Residual Bidirectional LSTM Model Using a Wearable Movement Monitoring Collar. *Agriculture* vol. 12 (2022).
99. Dickinson, E. R., Stephens, P. A., Marks, N. J., Wilson, R. P. & Scantlebury, D. M. Best practice for collar deployment of tri-axial accelerometers on a terrestrial quadruped to provide accurate measurement of body acceleration. *Anim. Biotelemetry* **8**, 1–8 (2020).
100. Wilson, R. P. *et al.* Luck in Food Finding Affects Individual Performance and Population Trajectories. *Curr. Biol.* **28**, 3871-3877.e5 (2018).
101. Sakamoto, K. Q. *et al.* Can Ethograms Be Automatically Generated Using Body Acceleration Data from Free-Ranging Birds? *PLoS One* **4**, e5379 (2009).
102. Viviant, M., Trites, A. W., Rosen, D. A. S., Monestiez, P. & Guinet, C. Prey capture attempts can be detected in Steller sea lions and other marine predators using accelerometers. *Polar Biol.* **33**, 713–719 (2010).
103. Watanabe, Y. Y. & Takahashi, A. Linking animal-borne video to accelerometers reveals prey capture variability. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2199–2204 (2013).
104. Kamminga, J. W. *et al.* Robust Sensor-Orientation-Independent Feature Selection for Animal Activity Recognition on Collar Tags. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2**, (2018).
105. Martiskainen, P. *et al.* Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. *Appl. Anim. Behav. Sci.* **119**, 32–38 (2009).
106. Kamminga, J. W., Janßen, L. M., Meratnia, N. & Havinga, P. J. M. Horsing around—A dataset comprising horse movement. *Data* **4**, 1–13 (2019).
107. Nweke, H. F., Teh, Y. W., Al-garadi, M. A. & Alo, U. R. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Syst. Appl.* **105**, 233–261 (2018).
108. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
109. Mao, A. *et al.* Cross-Modality Interaction Network for Equine Activity Recognition

Using Imbalanced Multi-Modal Data. *Sensors* vol. 21 (2021).

110. Kleanthous, N., Hussain, A., Khan, W., Sneddon, J. & Liatsis, P. Deep transfer learning in sheep activity recognition using accelerometer data. *Expert Syst. Appl.* **207**, 117925 (2022).
111. Eerdekens, A. *et al.* Resampling and Data Augmentation For Equines' Behaviour Classification Based on Wearable Sensor Accelerometer Data Using a Convolutional Neural Network. in *2020 International Conference on Omni-layer Intelligent Systems (COINS)* 1–6 (2020). doi:10.1109/COINS49042.2020.9191639.
112. Li, C. *et al.* Data Augmentation for Inertial Sensor Data in CNNs for Cattle Behavior Classification. *IEEE Sensors Lett.* **5**, 1–4 (2021).
113. Pan, Z., Chen, H., Zhong, W., Wang, A. & Zheng, C. A CNN-Based Animal Behavior Recognition Algorithm for Wearable Devices. *IEEE Sens. J.* **23**, 5156–5164 (2023).
114. Li, Z. & Ko, B. Naive semi-supervised deep learning using pseudo-label. 1358–1368 (2019).
115. Zhang, S., Li, Z., Yan, S., He, X. & Sun, J. Distribution Alignment: A Unified Framework for Long-tail Visual Recognition. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2361–2370 (2021). doi:10.1109/CVPR46437.2021.00239.
116. Gerych, W., Agu, E. & Rundensteiner, E. Classifying Depression in Imbalanced Datasets Using an Autoencoder- Based Anomaly Detection Approach. in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)* 124–127 (2019). doi:10.1109/ICOSC.2019.8665535.
117. Ruff, L. *et al.* Deep one-class classification. *35th Int. Conf. Mach. Learn. ICML 2018* **10**, 6981–6996 (2018).
118. Zhao, Z. *et al.* Improved Sensor-Based Animal Behavior Classification Performance through Conditional Generative Adversarial Network. (2022).
119. Chen, K. *et al.* Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Comput. Surv.* **54**, (2021).
120. Yurur, O., Liu, C. H. & Moreno, W. A survey of context-aware middleware designs for human activity recognition. *IEEE Commun. Mag.* **52**, 24–31 (2014).
121. Geng, C., Huang, S.-J. & Chen, S. Recent Advances in Open Set Recognition: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3614–3631 (2021).
122. Kamminga, J. W., Le, D. V & Havinga, P. J. M. Towards deep unsupervised representation learning from accelerometer time series for animal activity recognition. *Proc. 6th Work. Min. Learn. from Time Ser. - MileTS* (2020).
123. Yang, X. *et al.* Classification of broiler behaviours using triaxial accelerometer and machine learning. *Animal* **15**, 100269 (2021).
124. Mei, W. *et al.* Identification of aflatoxin-poisoned broilers based on accelerometer and machine learning. *Biosyst. Eng.* **227**, 107–116 (2023).
125. Cornou, C., Lundbye-Christensen, S. & Kristensen, A. R. Modelling and monitoring sows' activity types in farrowing house using acceleration data. *Comput. Electron. Agric.* **76**, 316–324 (2011).
126. Thompson, R., Matheson, S. M., Plötz, T., Edwards, S. A. & Kyriazakis, I. Porcine lie detectors: Automatic quantification of posture state and transitions in sows using inertial sensors. *Comput. Electron. Agric.* **127**, 521–530 (2016).
127. Papageorgiou, D., Rozen-Rechels, D., Nyaguthii, B. & Farine, D. R. Seasonality impacts collective movements in a wild group-living bird. *Mov. Ecol.* **9**, 1–12 (2021).
128. Paruelo, J. M., Epstein, H. E., Lauenroth, W. K. & Burke, I. C. ANPP estimates from NDVI for the central grassland region of the United States. *Ecology* **78**, 953–958 (1997).
129. Gebremeskel, G. *et al.* Droughts in East Africa: Causes, impacts and resilience. *Earth-*

Science Rev. **193**, 146–161 (2019).

130. Young, T. P., Stanton, M. L. & Christian, C. E. Effects of natural and simulated herbivory on spine lengths of Acacia drepanolobium in Kenya. *Oikos* **101**, 171–179 (2003).
131. Papageorgiou, D. *et al.* The multilevel society of a small-brained bird. *Curr. Biol.* **29**, R1120–R1121 (2019).
132. Guindre-Parker, S. & Rubenstein, D. R. Survival Benefits of Group Living in a Fluctuating Environment. *Am. Nat.* **195**, 1027–1036 (2020).
133. Muñoz Sabater, J. ERA5-Land monthly averaged data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (2019) doi:10.24381/cds.68d2bb30.
134. Huey, R. B. *et al.* Predicting organismal vulnerability to climate warming: roles of behaviour, physiology and adaptation. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 1665–1679 (2012).
135. Wong, B. B. M. & Candolin, U. Behavioral responses to changing environments. *Behav. Ecol.* **26**, 665–673 (2015).
136. Aureli, F. *et al.* Fission-Fusion Dynamics: New Research Frameworks. *Curr. Anthropol.* **49**, 627–654 (2008).
137. Couzin, I. D. Behavioral Ecology: Social Organization in Fission–Fusion Societies. *Curr. Biol.* **16**, R169–R171 (2006).
138. Dehnen, T. *et al.* Costs dictate strategic investment in dominance interactions. *Philos. Trans. R. Soc. London. Ser. B, Biol. Sci.* **377**, 20200447 (2022).
139. Johnson, N. K. Handbook of the Birds of the World, Volume 6. *Auk* **119**, 573–574 (2002).
140. Klarevas-Irby, J. A., Wikelski, M. & Farine, D. R. Efficient movement strategies mitigate the energetic cost of dispersal. *Ecol. Lett.* **24**, 1432–1442 (2021).
141. Hamilton, W. D. The genetical evolution of social behaviour. II. *J. Theor. Biol.* **7**, 17–52 (1964).
142. Rahman, A. *et al.* Cattle behaviour classification from collar, halter, and ear tag sensors. *Inf. Process. Agric.* **5**, 124–133 (2018).
143. Tamura, T. *et al.* Dairy cattle behavior classifications based on decision tree learning using 3-axis neck-mounted accelerometers. *Anim. Sci. J.* **90**, 589–596 (2019).
144. Riaboff, L. *et al.* Development of a methodological framework for a robust prediction of the main behaviours of dairy cows using a combination of machine learning algorithms on accelerometer data. *Comput. Electron. Agric.* **169**, 105179 (2020).
145. e-obs GmbH. E-Obs System Manual. 1–318 (2022).
146. Kays, R. *et al.* The Movebank system for studying global animal movement and demography. *Methods Ecol. Evol.* **13**, 419–431 (2022).
147. Berger, M. C. Movebank Acceleration Viewer. (2020).
148. Auer, E. *et al.* ELAN as flexible annotation framework for sound and image processing detectors. in *Seventh conference on International Language Resources and Evaluation [LREC 2010]* 890–893 (European Language Resources Association (ELRA), 2010).
149. Kranstauber, B., Smolla, M., Scharf, A. K. & Kranstauber, M. B. Package 'move'. (2023).
150. Wickham, H., Chang, W. & Wickham, M. H. Package 'ggplot2'. *Creat. elegant data Vis. using Gramm. Graph. Version* **2**, 1–189 (2016).
151. Allik, A. *et al.* Optimization of Physical Activity Recognition for Real-Time Wearable Systems: Effect of Window Length, Sampling Frequency and Number of Features. *Applied Sciences* vol. 9 (2019).
152. Banos, O., Galvez, J.-M., Damas, M., Pomares, H. & Rojas, I. Window Size Impact in Human Activity Recognition. *Sensors* vol. 14 6474–6499 (2014).

153. Lush, L. *et al.* Classification of sheep urination events using accelerometers to aid improved measurements of livestock contributions to nitrous oxide emissions. *Comput. Electron. Agric.* **150**, 170–177 (2018).
154. Arablouei, R. *et al.* In-situ classification of cattle behavior using accelerometry data. *Comput. Electron. Agric.* **183**, 106045 (2021).
155. Arablouei, R. *et al.* In-situ animal behavior classification using knowledge distillation and fixed-point quantization. *Smart Agric. Technol.* **4**, 100159 (2023).
156. Arablouei, R., Wang, Z., Bishop-Hurley, G. J. & Liu, J. Multimodal sensor data fusion for in-situ classification of animal behavior using accelerometry and GNSS data. *Smart Agric. Technol.* **4**, 100163 (2023).
157. Rao, R. B., Fung, G. & Rosales, R. On the dangers of cross-validation. An experimental evaluation. in *Proceedings of the 2008 SIAM international conference on data mining* 588–596 (SIAM, 2008).
158. Chollet, F. *Deep learning with Python*. (Simon and Schuster, 2021).
159. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
160. Abadi, M. *et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv Prepr. arXiv1603.04467* (2016).
161. Chollet, F. Xception: Deep learning with depthwise separable convolutions. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 1251–1258 (2017).
162. Wang, L., Arablouei, R., Alvarenga, F. A. P. & Bishop-Hurley, G. J. Animal Behavior Classification via Accelerometry Data and Recurrent Neural Networks. (2021).
163. Minati, L. *et al.* Accelerometer time series augmentation through externally driving a non-linear dynamical system. *Chaos, Solitons & Fractals* **168**, 113100 (2023).
164. Hassoun, M. H. *Fundamentals of artificial neural networks*. (MIT press, 1995).
165. Lee, S.-M., Yoon, S. M. & Cho, H. Human activity recognition from accelerometer data using Convolutional Neural Network. in *2017 ieee international conference on big data and smart computing (bigcomp)* 131–134 (IEEE, 2017).
166. Ha, S. & Choi, S. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. in *2016 international joint conference on neural networks (IJCNN)* 381–388 (IEEE, 2016).
167. Ignatov, A. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Appl. Soft Comput.* **62**, 915–922 (2018).
168. O'Shea, K. & Nash, R. An introduction to convolutional neural networks. *arXiv Prepr. arXiv1511.08458* (2015).
169. Liou, C.-Y., Cheng, W.-C., Liou, J.-W. & Liou, D.-R. Autoencoder for words. *Neurocomputing* **139**, 84–96 (2014).
170. Zhou, S., Xue, Z. & Du, P. Semisupervised stacked autoencoder with cotraining for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **57**, 3813–3826 (2019).
171. Bank, D., Koenigstein, N. & Giryes, R. Autoencoders. *Mach. Learn. data Sci. Handb. data Min. Knowl. Discov. Handb.* 353–374 (2023).
172. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conf. Comput. Vis. Pattern Recognit.* 770–778 (2015).
173. Wu, Z., Shen, C. & Van Den Hengel, A. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognit.* **90**, 119–133 (2019).
174. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.-A. Deep learning for time series classification: a review. *Data Min. Knowl. Discov.* **33**, 917–963 (2019).
175. Wang, Z., Yan, W. & Oates, T. Time series classification from scratch with deep

neural networks: A strong baseline. in *2017 International Joint Conference on Neural Networks (IJCNN)* 1578–1585 (2017). doi:10.1109/IJCNN.2017.7966039.

176. Tang, Y., Zhang, L., Teng, Q., Min, F. & Song, A. Triple Cross-Domain Attention on Human Activity Recognition Using Wearable Sensors. *IEEE Trans. Emerg. Top. Comput. Intell.* **6**, 1167–1176 (2022).
177. Hussain, A. *et al.* Long Short-Term Memory (LSTM)-Based Dog Activity Detection Using Accelerometer and Gyroscope. *Applied Sciences* vol. 12 (2022).
178. Medsker, L. R. & Jain, L. C. Recurrent neural networks. *Des. Appl.* **5**, 2 (2001).
179. Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **31**, 1235–1270 (2019).
180. Liseune, A., den Poel, D. Van, Hut, P. R., van Eerdenburg, F. J. C. M. & Hostens, M. Leveraging sequential information from multivariate behavioral sensor data to predict the moment of calving in dairy cattle using deep learning. *Comput. Electron. Agric.* **191**, 106566 (2021).
181. Kim, J. & Moon, N. Dog Behavior Recognition Based on Multimodal Data from a Camera and Wearable Device. *Applied Sciences* vol. 12 (2022).
182. Chambers, R. D. *et al.* Deep Learning Classification of Canine Behavior Using a Single Collar-Mounted Accelerometer: Real-World Validation. *Animals* vol. 11 (2021).
183. Mutegeki, R. & Han, D. S. A CNN-LSTM approach to human activity recognition. in *2020 international conference on artificial intelligence in information and communication (ICAIIC)* 362–366 (IEEE, 2020).
184. Xia, K., Huang, J. & Wang, H. LSTM-CNN Architecture for Human Activity Recognition. *IEEE Access* **8**, 56855–56866 (2020).
185. Wang, L., Arablouei, R., Alvarenga, F. A. P. & Bishop-Hurley, G. J. Classifying animal behavior from accelerometry data via recurrent neural networks. *Comput. Electron. Agric.* **206**, 107647 (2023).
186. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (MIT press, 2016).
187. Molnar, C., Casalicchio, G. & Bischl, B. Interpretable machine learning–a brief history, state-of-the-art and challenges. in *Joint European conference on machine learning and knowledge discovery in databases* 417–431 (Springer, 2020).
188. Zweig, M. H. & Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **39**, 561–577 (1993).
189. Stehman, S. V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* **62**, 77–89 (1997).
190. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA)-Protein Struct.* **405**, 442–451 (1975).
191. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157 (1947).
192. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* vol. 57 10–25080 (Austin, TX, 2010).
193. Welch, B. L. The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika* **34**, 28–35 (1947).
194. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
195. Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**, 1–12 (2004).
196. Shavelson, R. J., Baxter, G. P. & Gao, X. Sampling variability of performance assessments. *J. Educ. Meas.* **30**, 215–232 (1993).
197. Cowlishaw, G. C. *Trade-offs between feeding competition and predation risk in baboons*. (University of London, University College London (United Kingdom), 1994).

198. Karnieli, A. *et al.* Use of NDVI and land surface temperature for drought assessment: Merits and limitations. *J. Clim.* **23**, 618–633 (2010).

199. Zhang, Q., Kong, D., Singh, V. P. & Shi, P. Response of vegetation to different time-scales drought across China: Spatiotemporal patterns, causes and implications. *Glob. Planet. Change* **152**, 1–11 (2017).

200. Vicente-Serrano, S. M. *et al.* Response of vegetation to drought time-scales across global land biomes. *Proc. Natl. Acad. Sci.* **110**, 52–57 (2013).

201. Iverson, A. R., Humple, D. L., Cormier, R. L. & Hull, J. Land cover and NDVI are important predictors in habitat selection along migration for the Golden-crowned Sparrow, a temperate-zone migrating songbird. *Mov. Ecol.* **11**, 1–19 (2023).

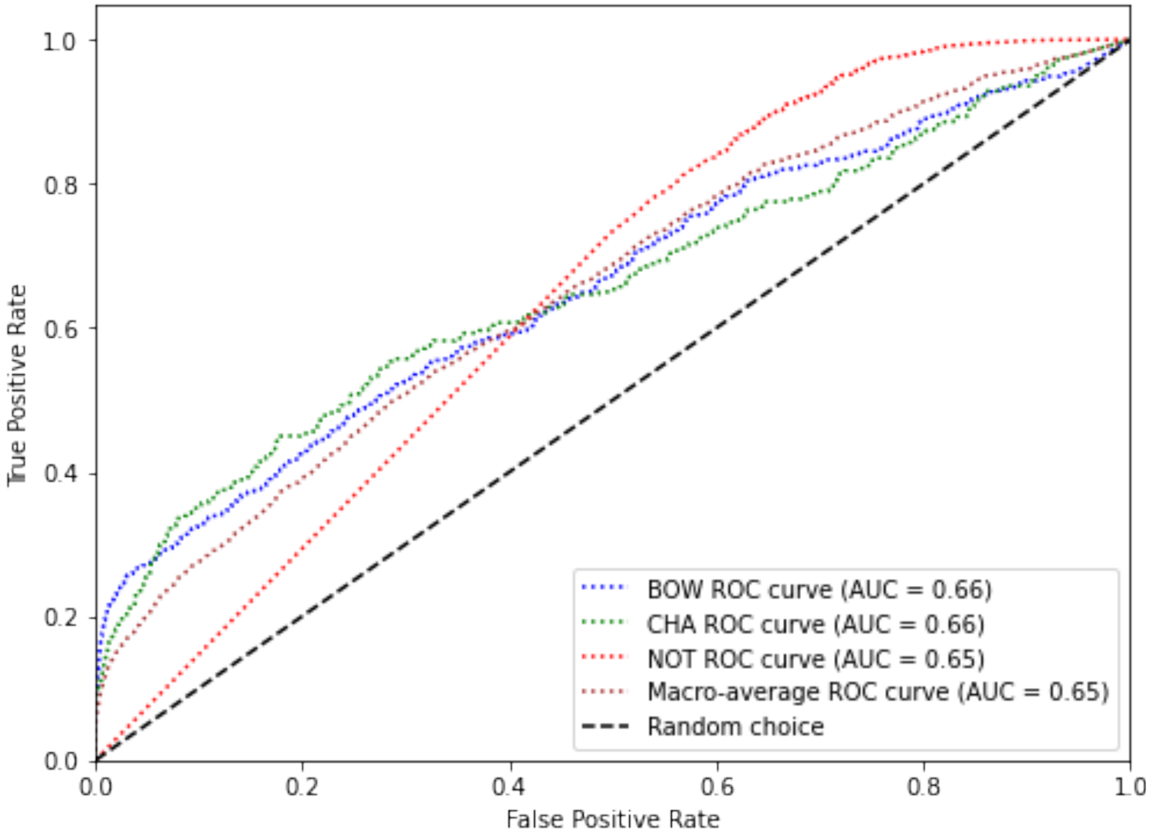
202. Fattorini, N. *et al.* Animal conflicts escalate in a warmer world. *Sci. Total Environ.* **871**, (2023).

203. McWilliams, K. M., Sandler, A. G., Atkins, G. J., Henson, S. M. & Hayward, J. L. Courtship and copulation in Glaucous-winged Gulls, <em>Larus glaucescens</em>, and the influence of environmental variables. *Wilson J. Ornithol.* **130**, 270–285 (2018).

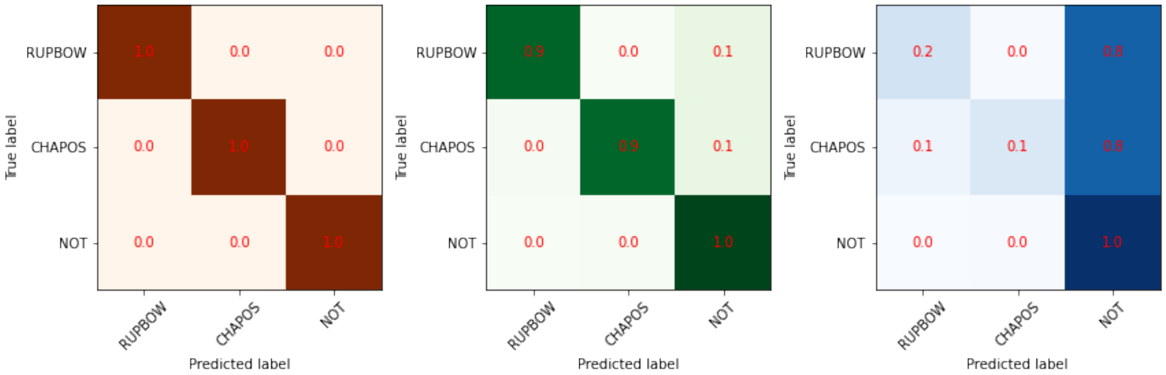
Appendices

Appendix I: MLP visual metrics for 1 second window size across all cross-validation folds

ROC curve

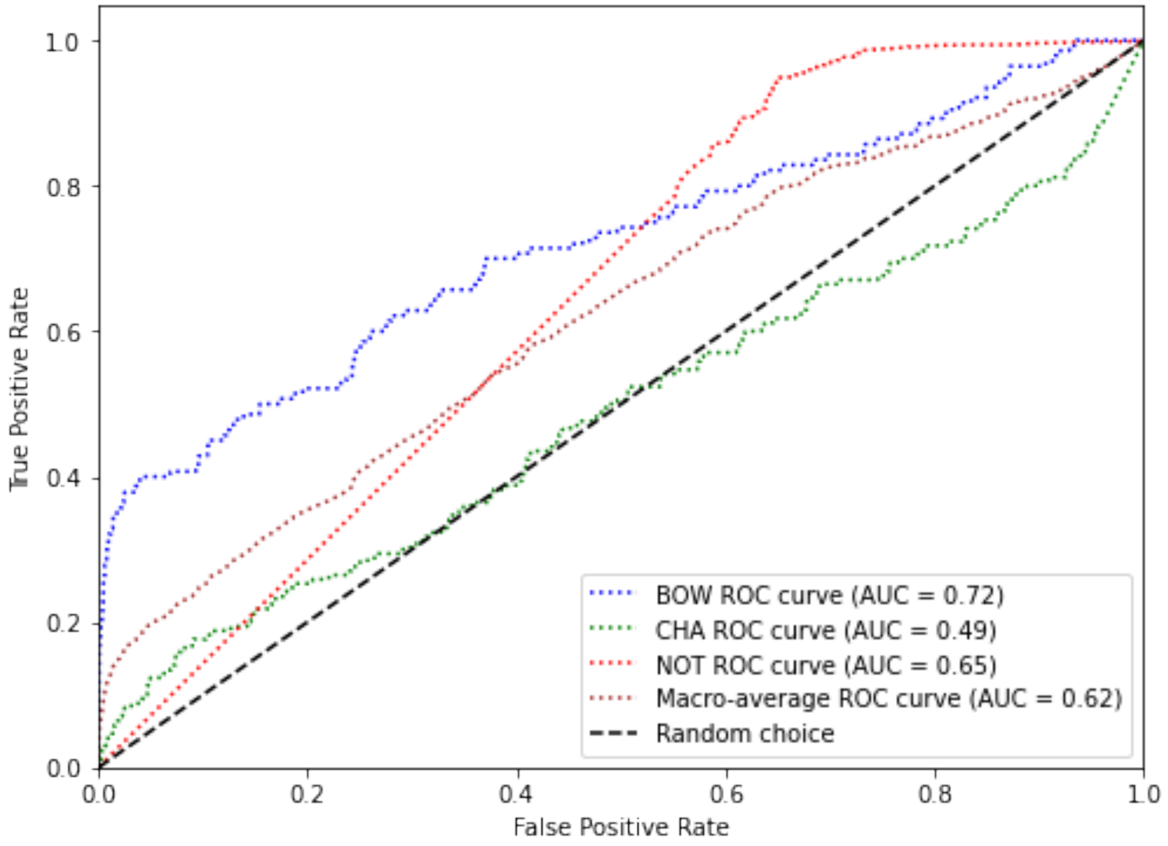


Confusion matrices for train, validation, and test set

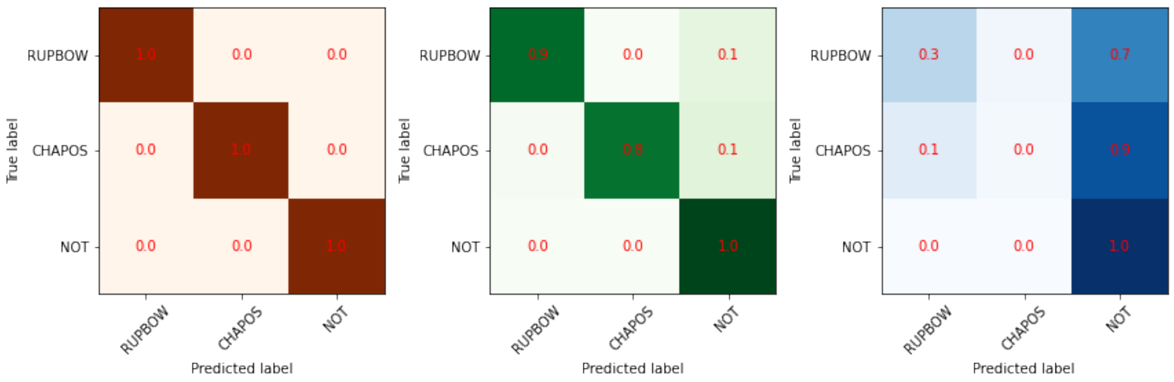


Appendix II: MLP visual metrics for 3 seconds window size for all cross-validation folds

ROC curve

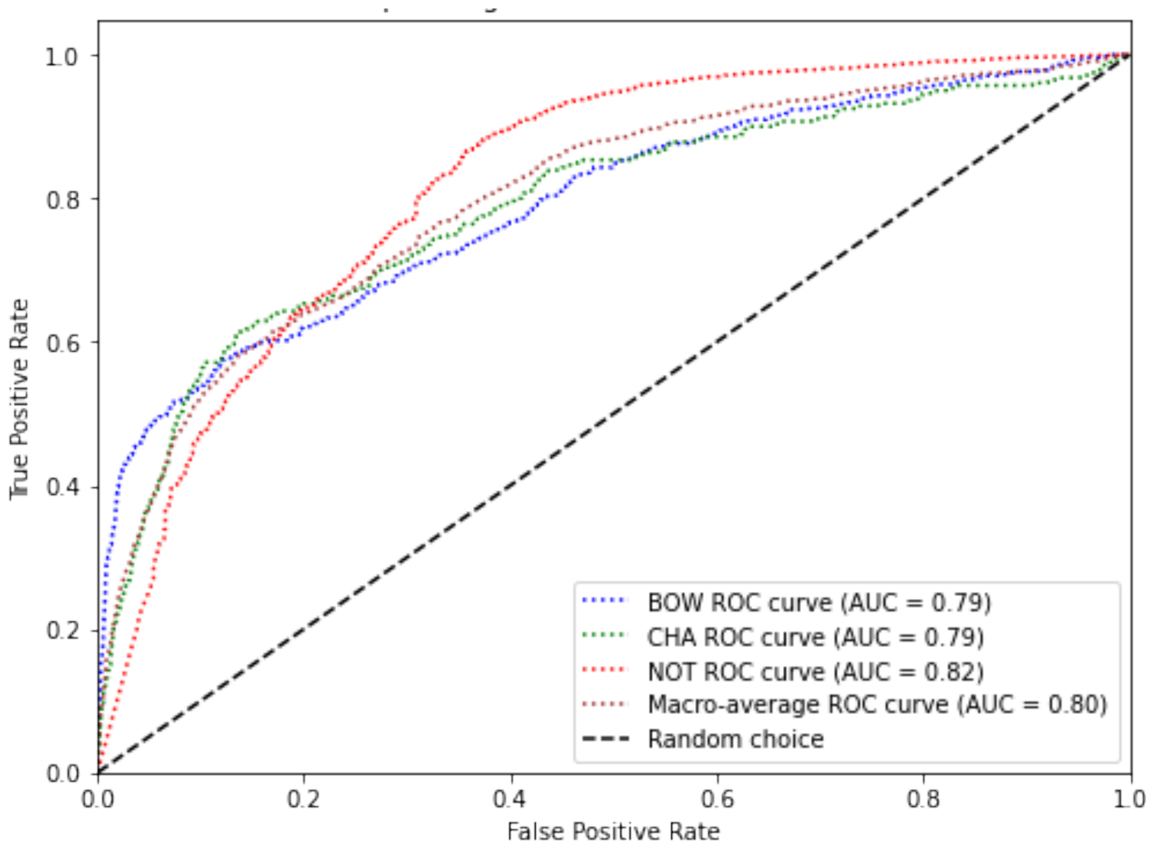


Confusion matrices for train, validation, and test set

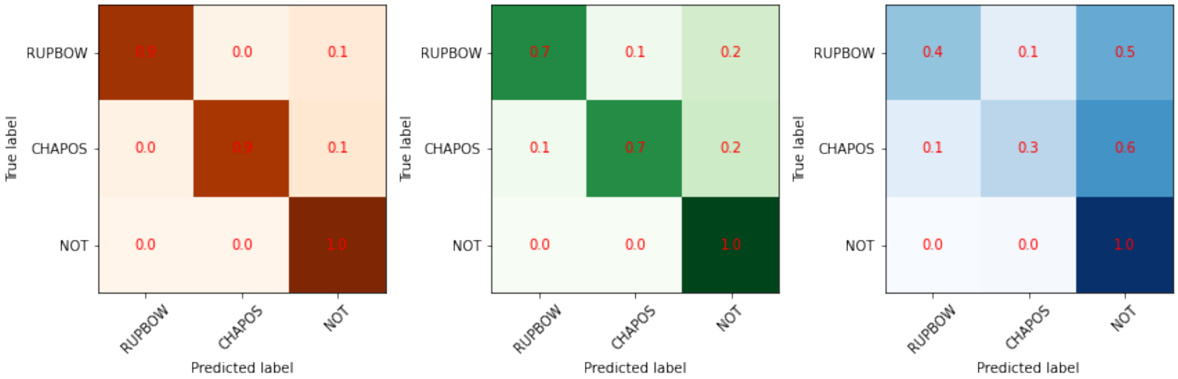


Appendix III: CNN visual metrics for 1 second window size for all cross-validation folds

ROC curve

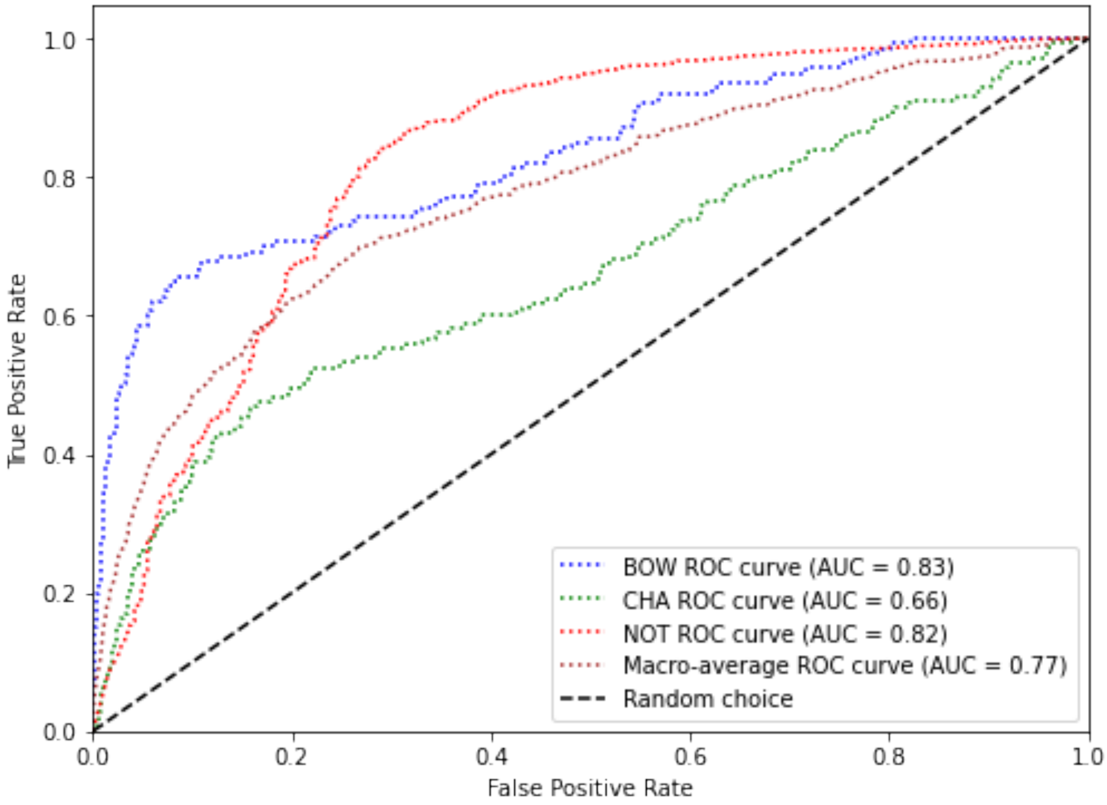


Confusion matrices for train, validation, and test set

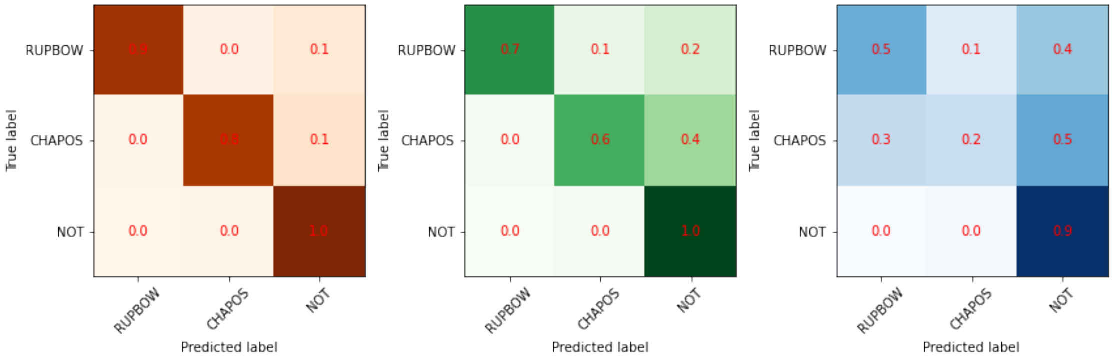


Appendix IV: CNN visual metrics for 3 seconds window size for all cross-validation folds

ROC curve

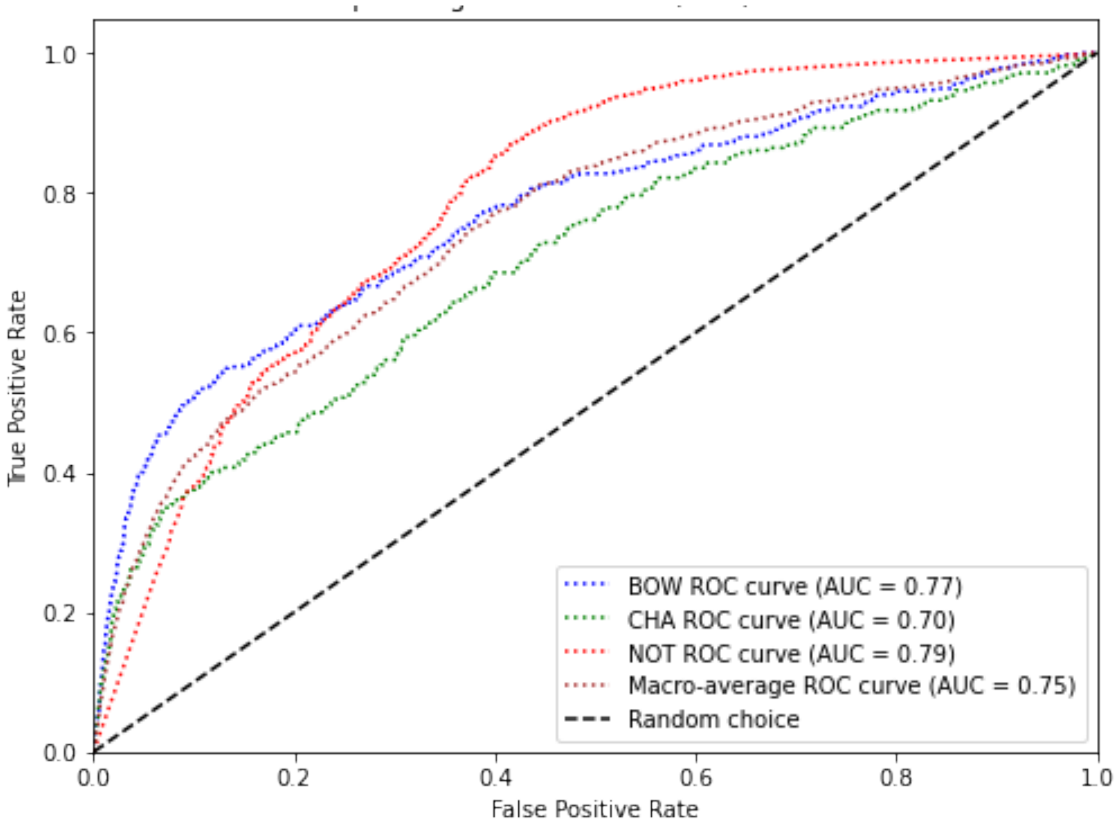


Confusion matrices for train, validation, and test set

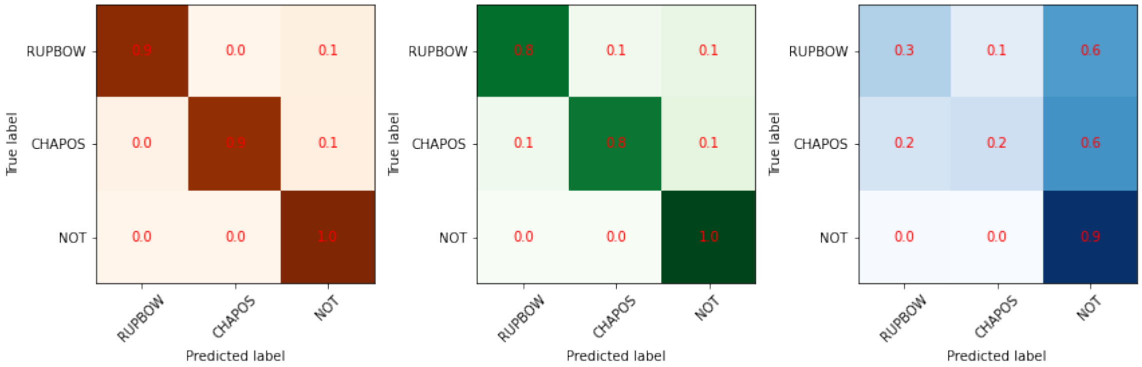


Appendix V: TCDA-CNN visual metrics for 1 second window size for all cross- validation folds

ROC curve

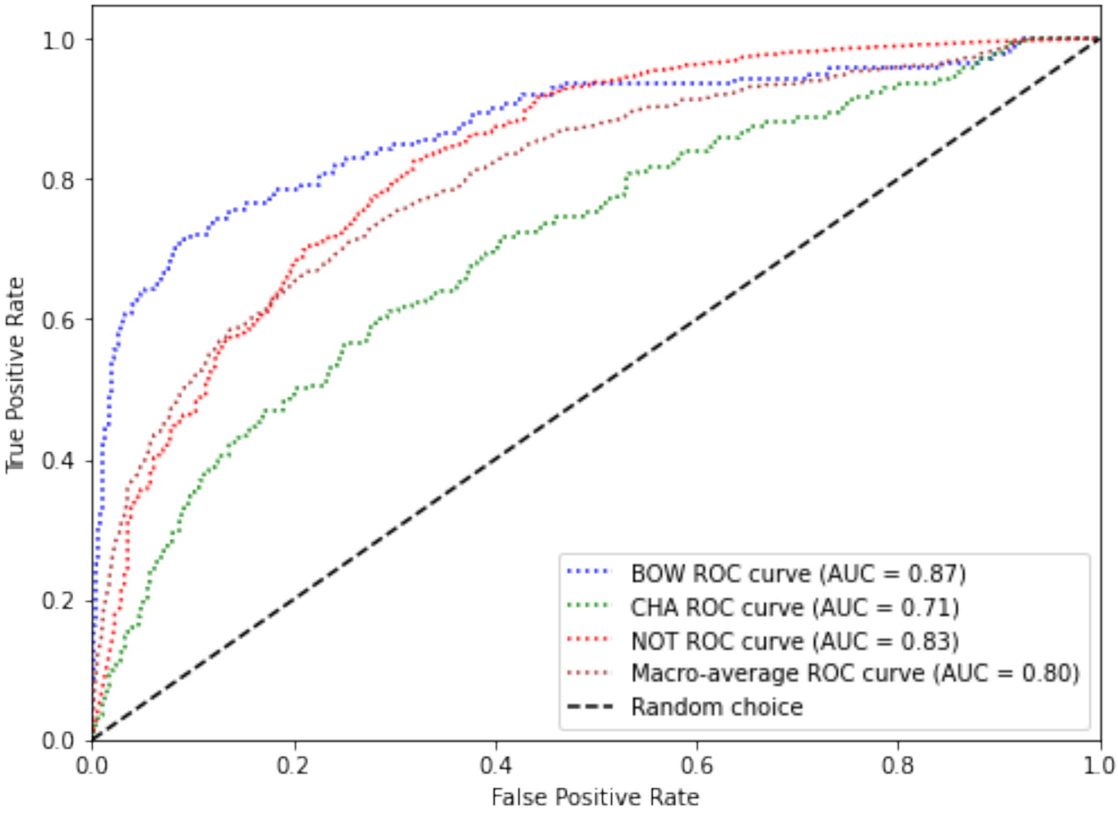


Confusion matrices for train, validation, and test set

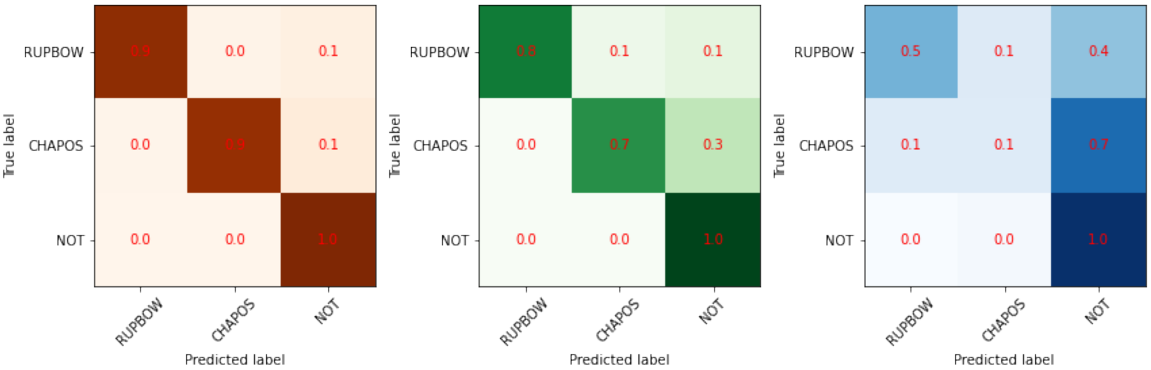


Appendix VI: TCDA-CNN visual metrics for 3 seconds window size for all cross- validation folds

ROC curve

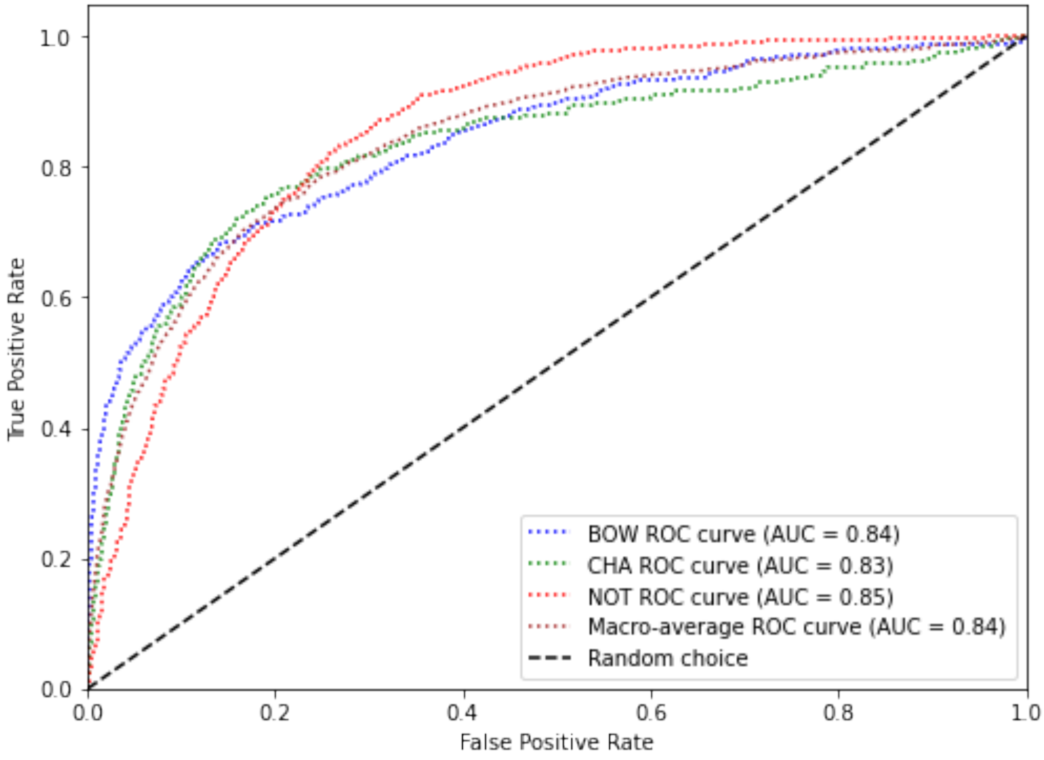


Confusion matrices for train, validation, and test set

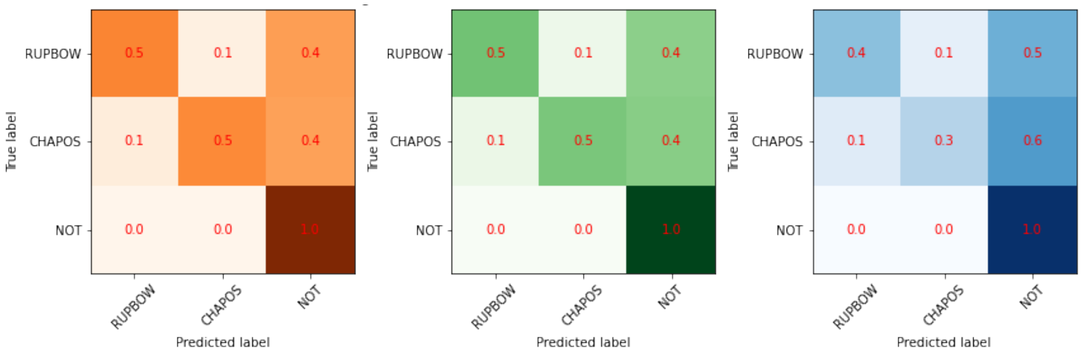


Appendix VII: Autoencoder visual metrics for 1 second window size for all cross- validation folds

ROC curve

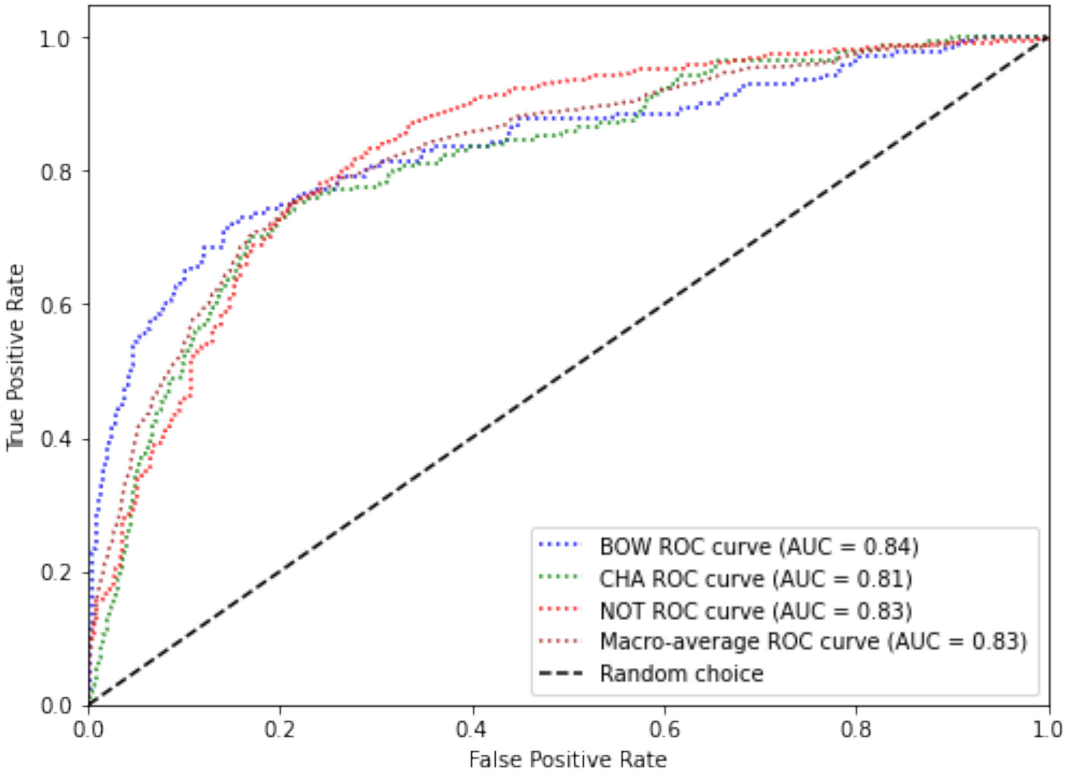


Confusion matrices for train, validation, and test set

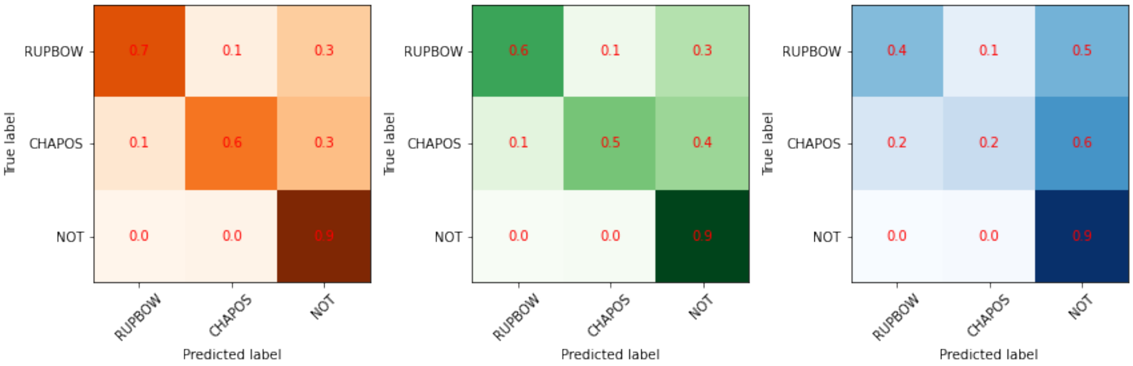


Appendix VIII: Autoencoder visual metrics for 3 seconds window size for all cross-validation folds

ROC curve

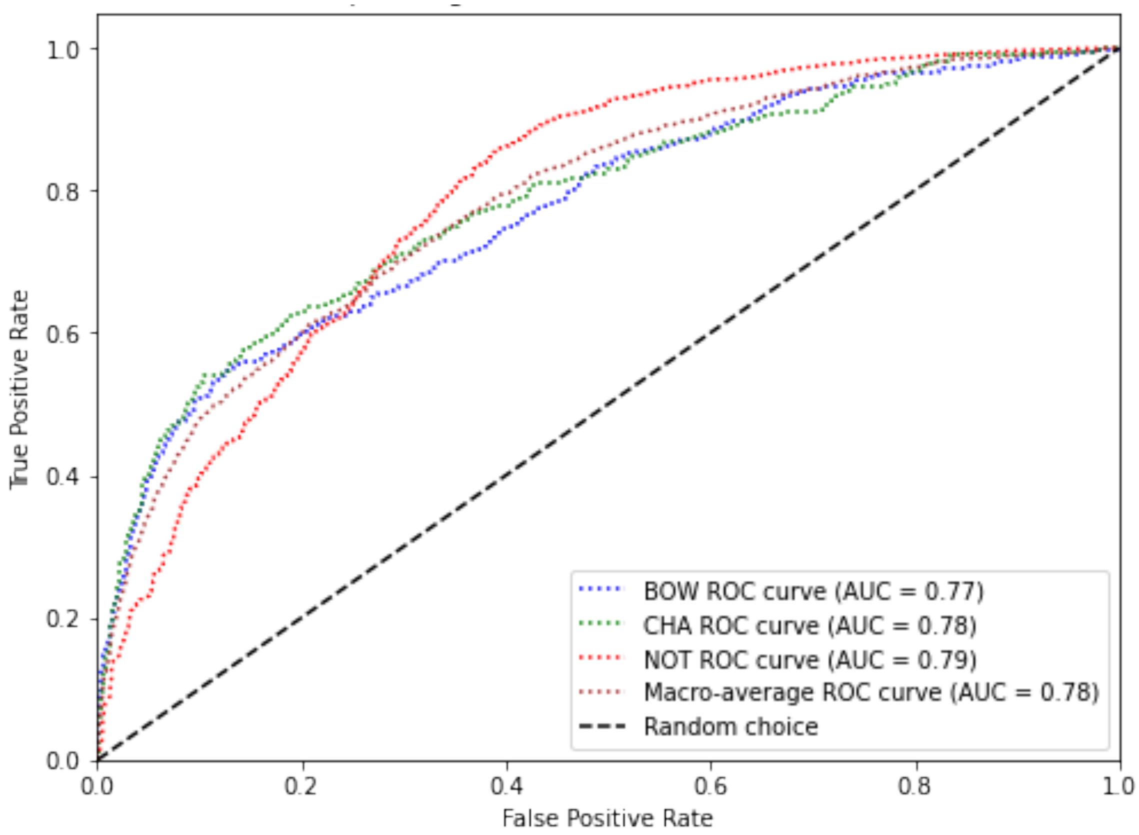


Confusion matrices for train, validation, and test set

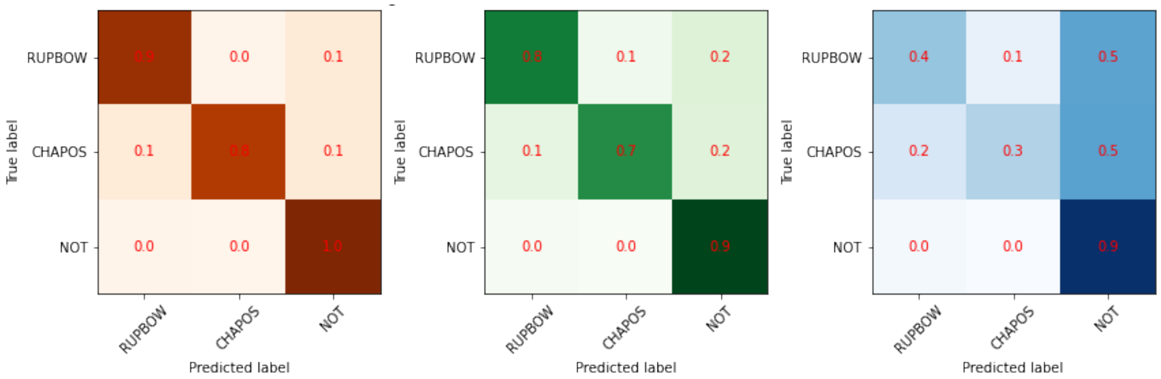


Appendix IX: ResNet Visual Metrics for 1 second window size for all cross- validation folds

ROC curve

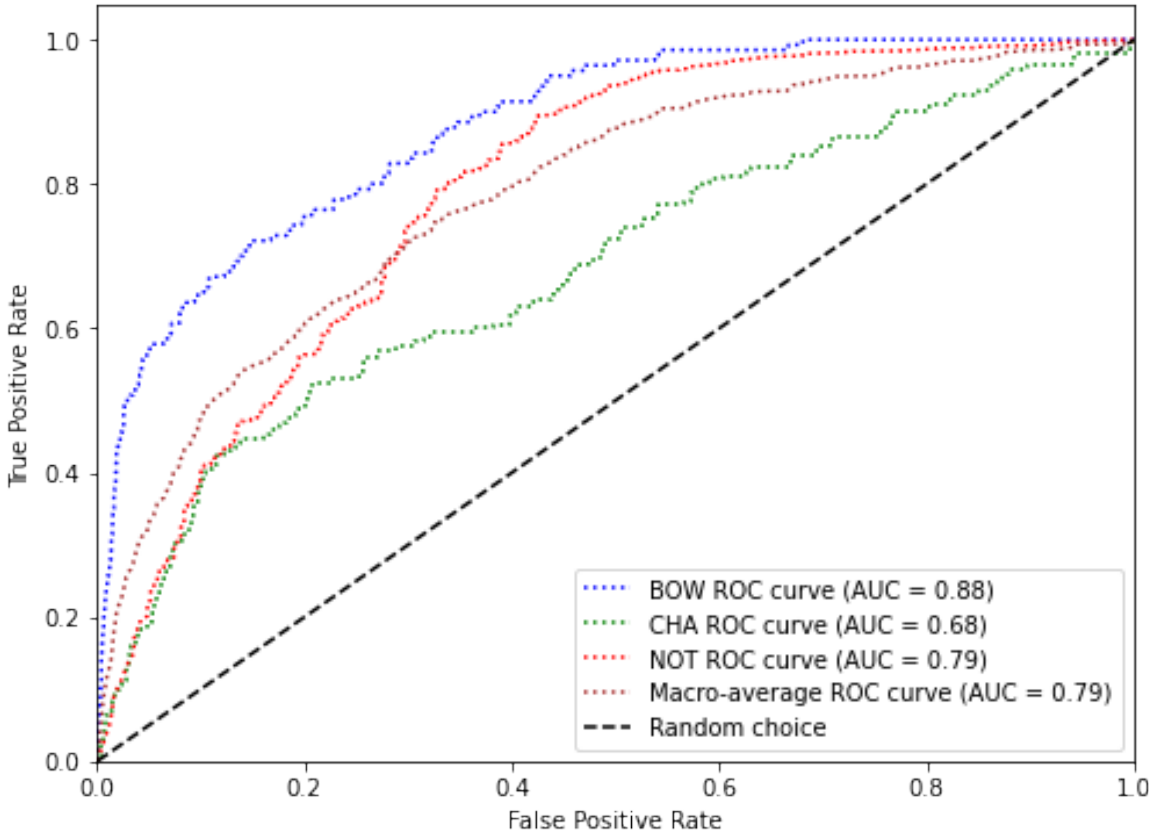


Confusion matrices for train, validation, and test set

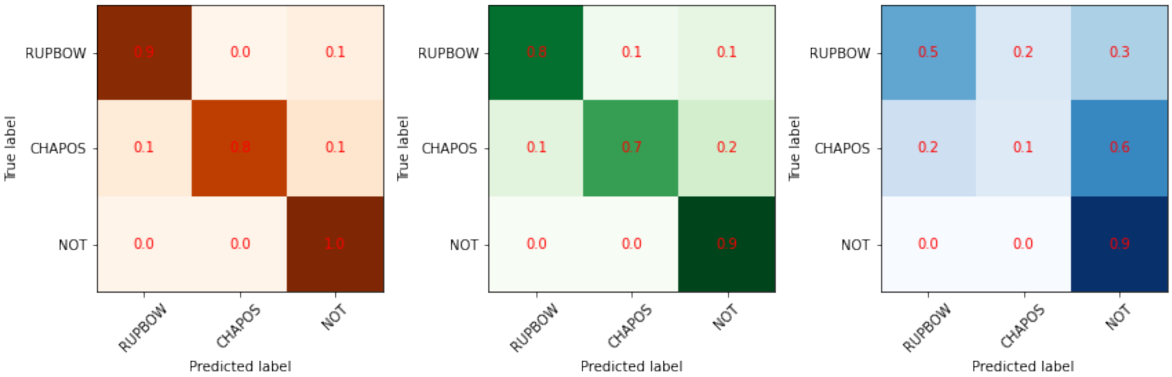


Appendix X: ResNet Visual Metrics for 3 seconds window size for all cross- validation folds

ROC curve

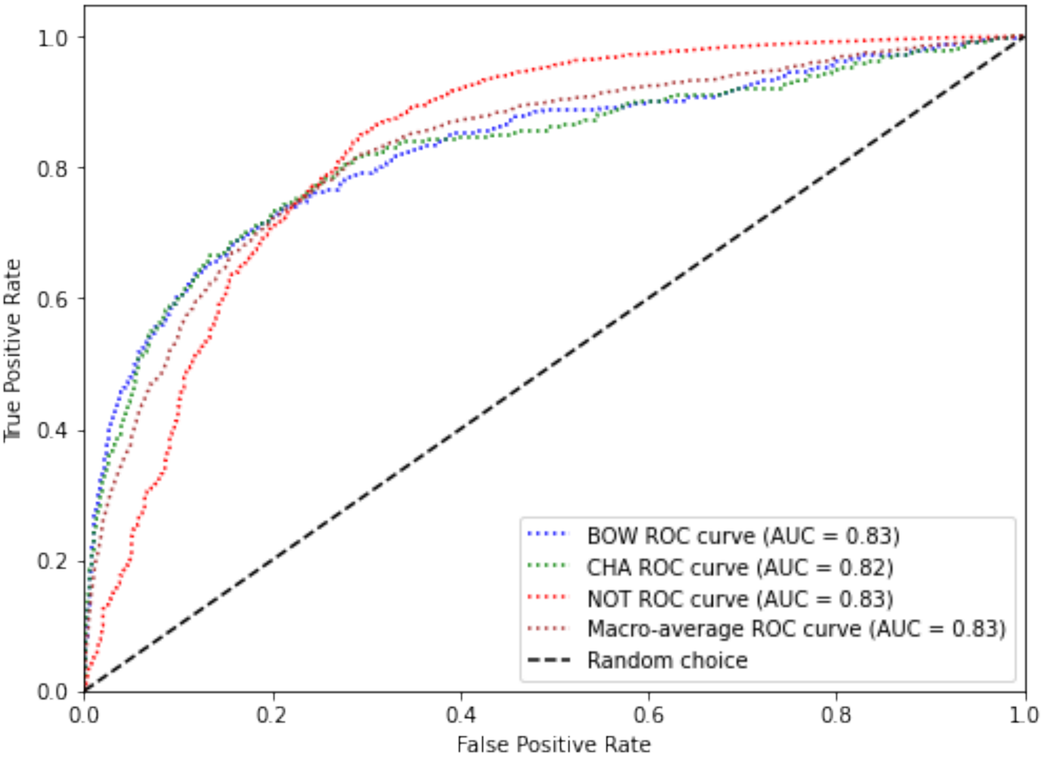


Confusion matrices for train, validation, and test set

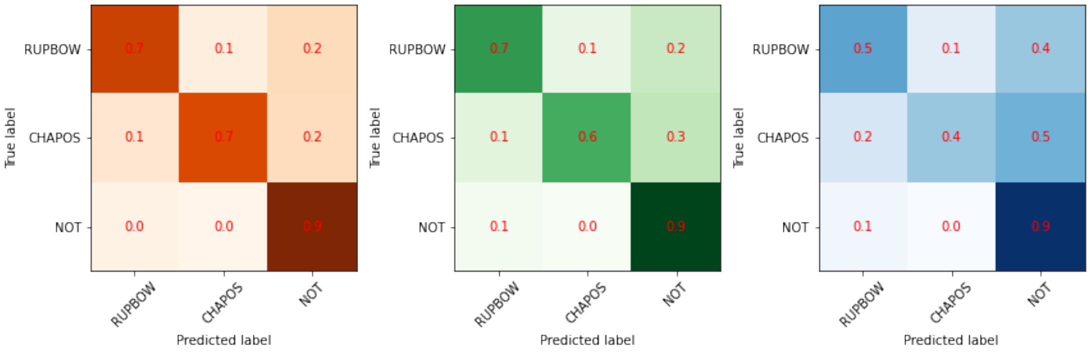


Appendix XI: CNN-LSTM Visual Metrics for 1 second window size for all cross- validation folds

ROC curve

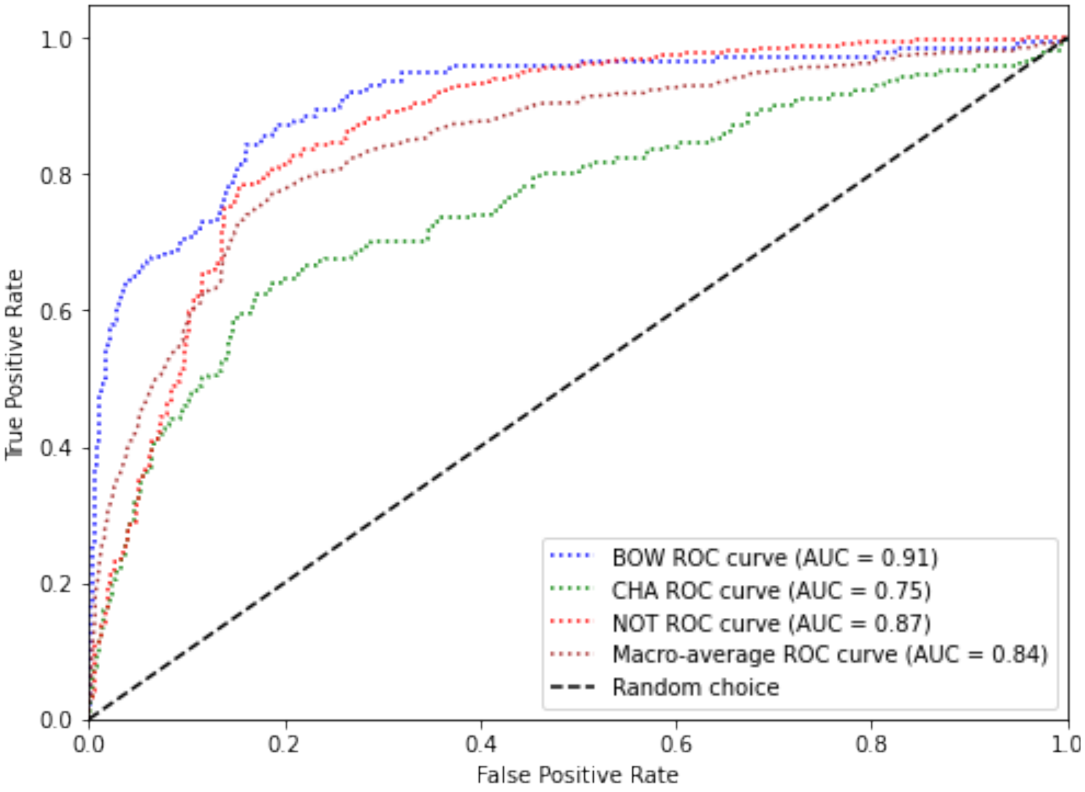


Confusion matrices for train, validation, and test set

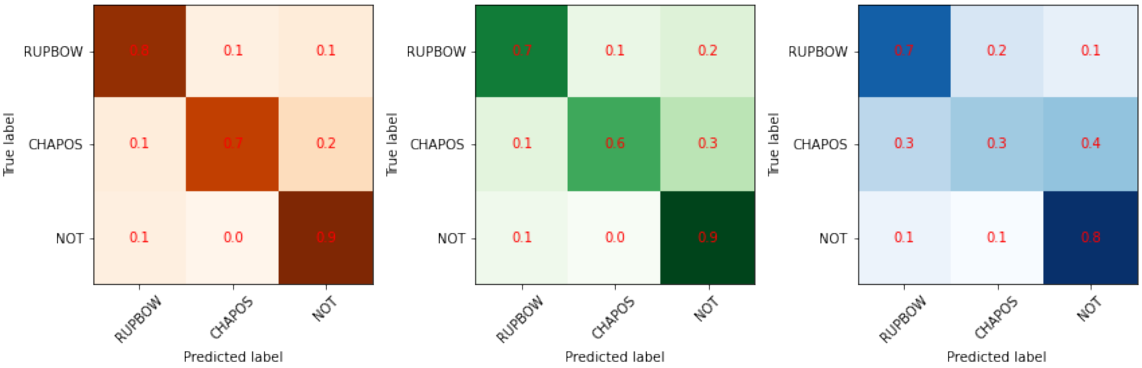


Appendix XII: CNN-LSTM Visual Metrics for 3 seconds window size for all cross-validation folds

ROC curve

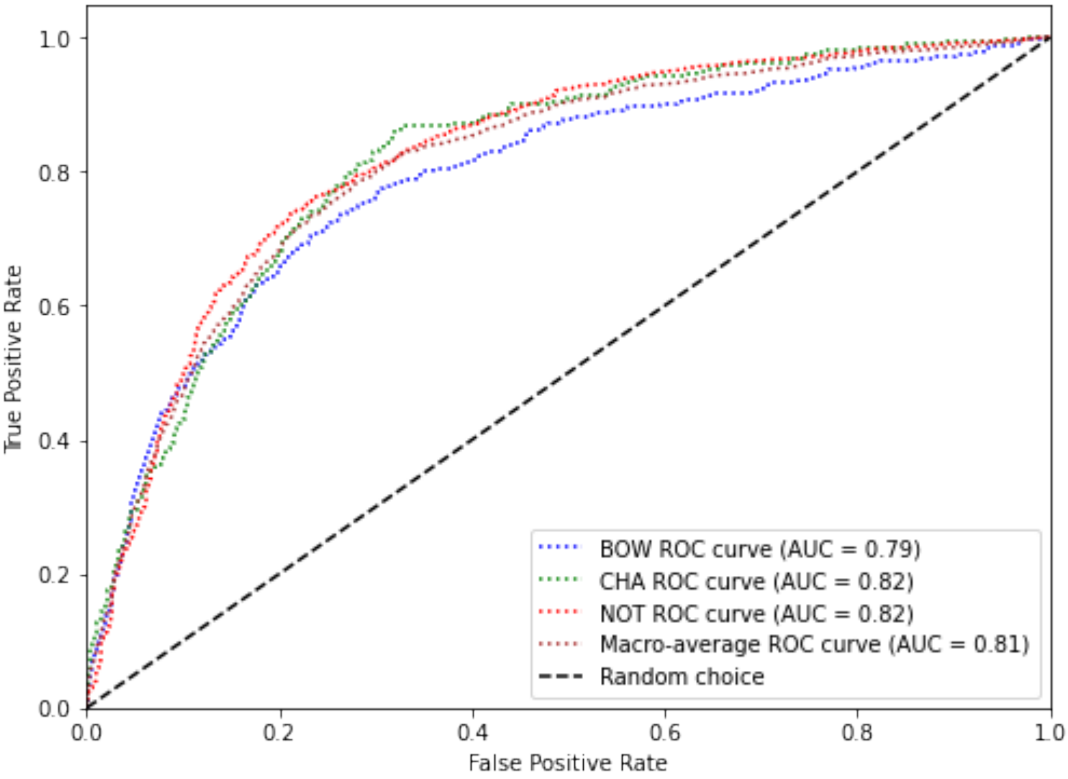


Confusion matrices for train, validation, and test set

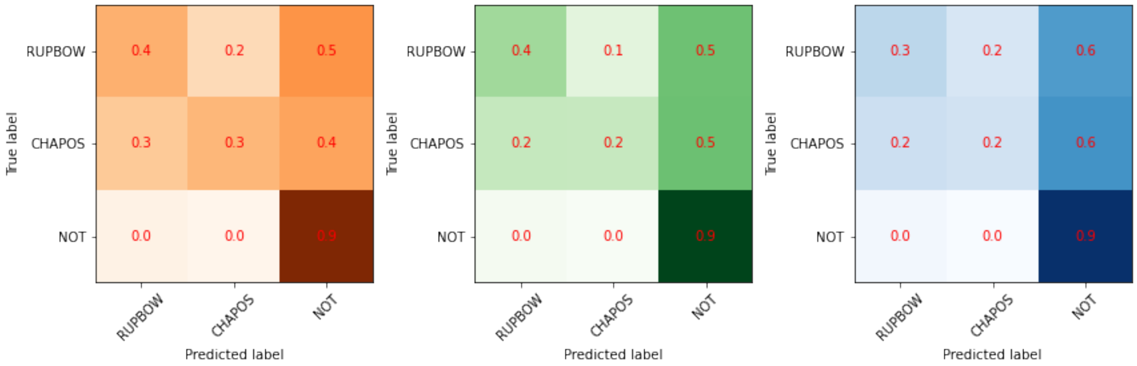


Appendix XIII: LSTM-CNN Visual Metrics for 1 second window size for all cross-validation folds

ROC curve

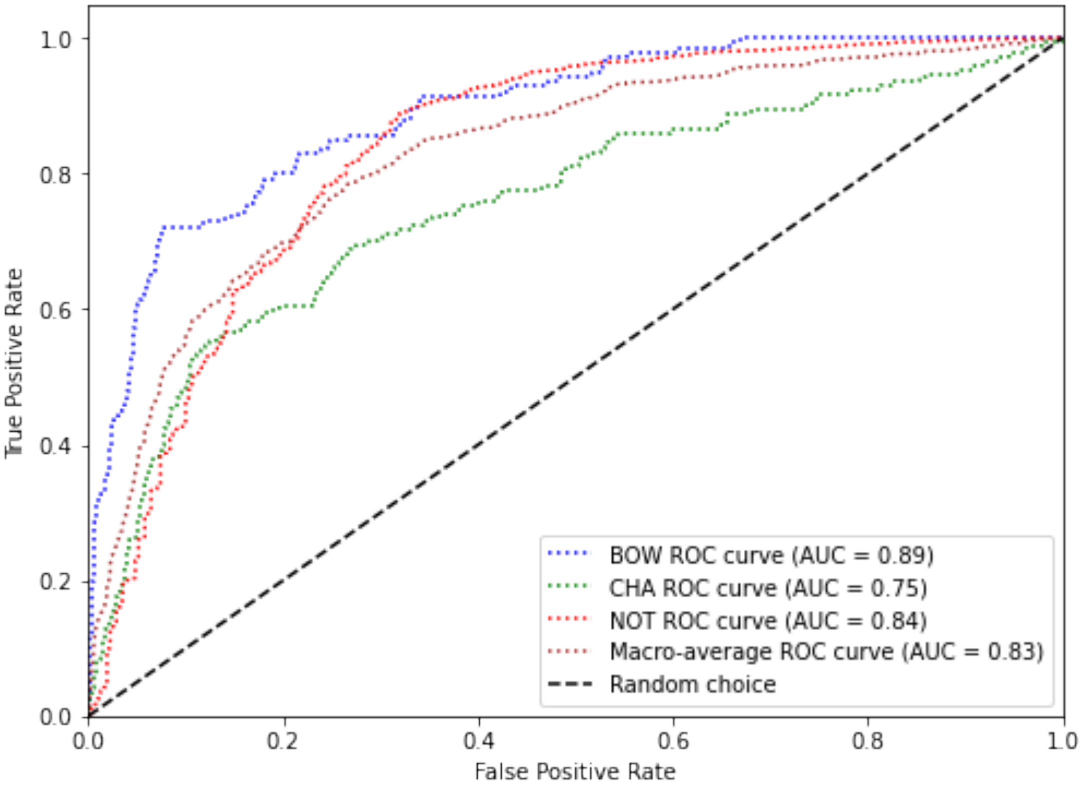


Confusion matrices for train, validation, and test set

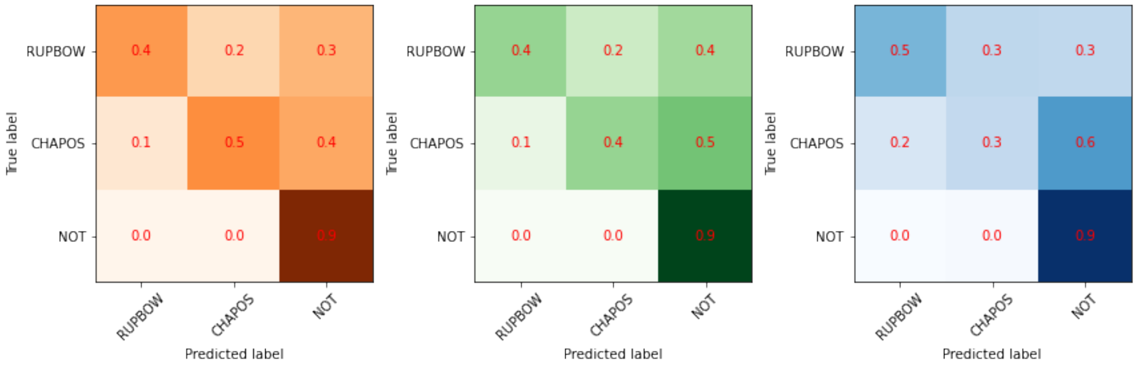


Appendix XI: LSTM-CNN Visual Metrics for 3 seconds window size for all cross-validation folds

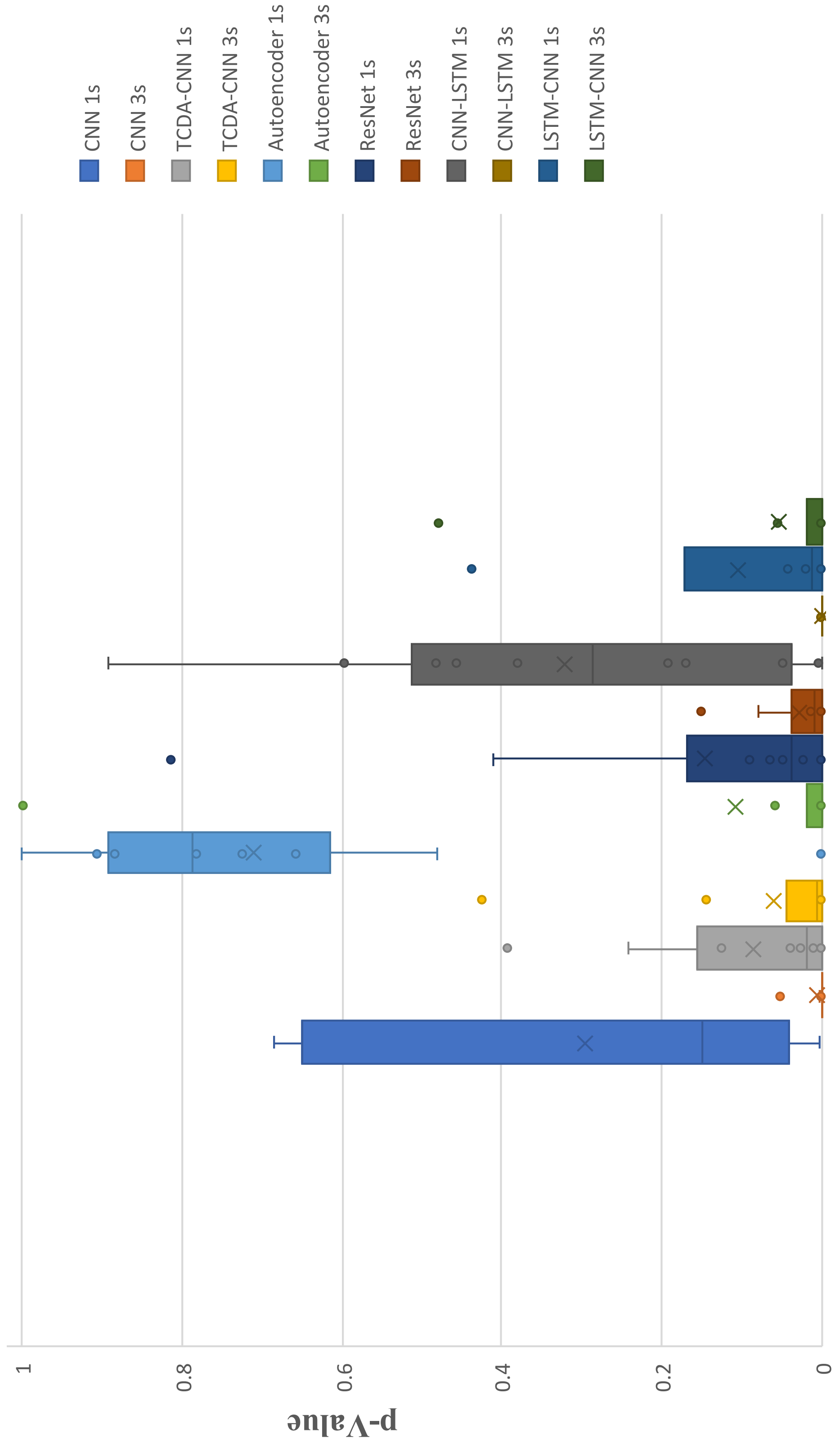
ROC curve



Confusion matrices for train, validation, and test set



Appendix XV: McNemar’s test results for each algorithm and window size



Statement of Authorship

I hereby declare that I have used no other sources and assistance other than those indicated and cited. All passages quoted from publications or paraphrased from these sources are indicated as such. This thesis was not submitted in any form for another degree or diploma at any university or other institution.

Zurich, 29.02.2024