On Computational Models of Animal Movement Behaviour

Kehinde Owoeye

Department of Computer Science

University College London

Thesis submitted in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

of

University College London

July 5, 2021

Declaration

I, Kehinde Owoeye, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

_____

Kehinde Owoeye

Abstract

Finding structures and patterns in animal movement data is essential towards understanding a variety of behavioural phenomena, as well as shedding light into the relationships between animals among conspecifics and across different taxa with respect to their environments. The recent advances in the field of computational intelligence coupled with the proliferation of low-cost telemetry devices have made the gathering and analyses of behavioural data of animals in their natural habitat and in a wide range of context possible with aid of devices such as Global Positioning System (GPS).

The sensory input that animals receive from their environment, and the corresponding motor output, as well as the neural basis of this relationship most especially as it affects movement, encode a lot of information regarding the welfare and survival of these animals and other organisms in nature's ecosystem. This has huge implications in the area of biodiversity monitoring, global health and understanding disease progression. Encoding, decoding and quantifying these functional relationships however can be challenging, boring and labour intensive. Artificial intelligence holds promise in solving some of these problems and even stand to benefit as understanding natural intelligence for instance can aid in the advancement of artificial intelligence.

In this thesis, I investigate and propose several computational methods leveraging information theoretic metrics and also modern machine learning methods including supervised, unsupervised and a novel combination of both towards understanding, predicting, forecasting and quantifying a variety of animal movement phenomena at different time scales across different taxa and species. Most importantly the models proposed in this thesis tackle important problems bordering on human and animal welfare as well as their intersection. Crucially, I investigate several information theoretic metrics towards mining animal movement data, after which I propose machine learning and statistical techniques for automatically quantifying abnormal movement behaviour in sheep with Batten disease using unsupervised methods. In addition, I propose a predictive model capable of forecasting migration patterns in Turkey vulture as well as their stop-over decisions using bidirectional recurrent neural networks. And finally, I propose a model of sheep movement behaviour in a flock leveraging insights in cognitive neuroscience with modern deep learning models.

Overall, the models of animal movement behaviour developed in this thesis are useful to a wide range of scientists in the field of neuroscience, ethology, veterinary science, conservation and public health. Although these models have been designed for understanding and predicting animal movement behaviour, in a lot of cases they scale easily into other domains such as human behaviour modelling with little modifications. I

highlight the importance of continuous research in developing computational models of animal movement behaviour towards improving our understanding of nature in relation to the interaction between animals and their environments.

Impact Statement

The advances in the field of computational intelligence and telemetry technologies have made the gathering of behavioural data from a variety of sources including animals in their natural habitats easier and their analyses feasible. This has made the study of animals under different conditions, habitats and scales possible thus shedding more light into their behaviour with applications in understanding their decision making processes, understanding factors governing their migration time, route and stopovers, tracking and monitoring their welfare and not limited to the understanding of how environmental variables influence their behaviour.

In this thesis, I have addressed a variety of problems that have potential impact across a wide range of disciplines involved in animal movement behaviour studies. The problems addressed in this thesis in addition to the results and analyses carried out have implications for researchers in the field of animal movement behaviour including ethologists, neuroscientists, veterinary physicians, conservationists as well as public health practitioners. Outside of animal behaviour, a considerable portion of the methods proposed in this thesis can also be used to model human behaviour. In general, I have developed computational techniques that often scale beyond the exact animal/species used in this thesis and are useful in modelling the behaviour of other animals with some characteristics similar to the ones considered in this thesis.

In chapter three of this work, I have investigated several information theoretic metrics towards mining animal movement data while in chapter four, I have come up with automatic machine learning methods for finding and quantifying abnormality in sheep movement patterns. These methods altogether are useful for poultry farmers and veterinary physicians in automatically monitoring real time the health and welfare of animals in their care. As a consequence, they tend to benefit from improved diagnosis, reduced labour and improved efficiency in their daily work. In addition, it is important for assessing the efficacy of therapeutic interventions in preclinical trials during drug discovery and development.

In chapter five of this work, the method proposed in forecasting migration patterns in Turkey vulture can help address biodiversity and conservation problems which is a core component of the united nations (UN) sustainable development goals. Outside of animal behaviour, it can be used in forecasting the onset of a disease for healthcare applications given some physiological measurements.

In chapter six, I have proposed methods with the potential to help address some of the global health risks such as avian flu and other zoonotic infections spread by migratory birds. Outside of this application, the techniques proposed can help address data bias issues for classification tasks in light of recent efforts at

reducing the biases in machine learning models.

Finally, the approach proposed in building predictive models of sheep movement behaviour in chapter seven for example has implications in understanding the neural basis underlying decision making processes of sheep and other animals potentially given the right set of longitudinal measurements. Beyond the study of animal behaviour, this approach can be used in other fields where understanding and prediction of behaviour is becoming increasingly important such as sport analytics, virtual assistants and in assisted living technologies.

The results and findings of this thesis have either been published or under review in journals, conference and workshop venues thus making them accessible to researchers and policy makers who will potentially benefit from the findings.

Acknowledgements

Special acknowledgement goes to my sponsor the Petroleum Technology Development Fund (PTDF) for their financial support. I am also grateful to the National University Commission (NUC) Nigeria for their role in the award of the scholarship. Also to the former president of Nigeria Goodluck Jonathan GCFR who initiated this scheme towards training the next generation of researchers in critical areas necessary for national develpment, thank you.

Special thanks to my supervisors Prof. Steve Hailes and Prof. Mirco Musolesi who have been supportive and helpful in the past four years. I also want to thank staff, both teaching and non-teaching in the UCL Computer Science department who have been helpful in the last four years.

My PhD journey would have been more stressful without my colleagues at the basement in Gower Street, I appreciate you all for the wonderful times. Also to my friends Taiwo, Tosin, Joseph and others too numerous to mention, thank you all for your support.

Finally, I would like to thank my family for their support, prayers and encouragement, without which this journey wouldn't have been successful. And to everyone else that have not been mentioned here, I would like to express my heartfelt appreciation to you all.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In this chapter, the aims and objectives of this thesis are addressed as well as the thesis structure.

Movement behaviour data of animals and by extension humans contain a wealth of information for understanding a wide range of phenomena. This is useful for example in answering important biological questions associated but not limited to migration, feeding, reproduction, social relationships, conservation, decision making and general physiological wellbeing. This dataset of movement behaviour is even of more paramount interest to scientists in the field of neuroscience, conservation, neuro-ethology, zoology and veterinary science for a variety of reasons. While a neuroscientist wants to understand the neural basis underlying animal movement behaviour, a conservationist is curious for example to understand how climate change has affected breeding or stopover sites of migrating birds.

Due to technological advances in manufacturing low cost miniaturised sensors, modern tracking devices such as Global Positioning System (GPS) and accelerometer are becoming ubiquitous for gathering movement behavioural data at varying resolutions and spatiotemporal scales across continents. This makes it possible for animal behaviour researchers to answer previously unanswered important and challenging questions. Analysing, modelling and obtaining useful insights from the vast amount of movement data collected from these sensors however can be a daunting and challenging task due to the limitations arising from the data collection process itself such as missing data due to loggers failure, non-stationarity, biasness, multimodal nature of behaviour data, presence of noise and most importantly due to the limitations of popular classical

statistical tools. Developing computational methods leveraging state of the art statistical and machine learning methods is therefore essential so as to be able to obtain the necessary utility from this data, answer new and previously unanswered biological questions in addition to generating new hypotheses.

Given the rapid and continuous advances in the field of machine learning and artificial intelligence, new capabilities for understanding and obtaining insights from data are spurring new wave of discoveries across a wide range of scientific endeavours. Techniques such as supervised, semi-supervised, unsupervised and reinforcement learning models are finding useful applications to scientists in the biological sciences for example in drug discovery [1], mammography [2], protein folding [3], disease progression [4]. While advances in these biological fields appear to be progressing at a fast pace with the aid of these techniques, their application in the field of animal behaviour is slow most especially in the field of animal movement behaviour. In this work, I leverage these advances in computational intelligence to build models of animal movement behaviour. These models find applications across a variety of animal movement behaviour related fields details of which can be found below.

1.1 Objectives

The essence of this thesis is to develop computational methods (leveraging machine learning models) useful in modelling the movement behaviour of animals in a wide range of context. The models proposed here have been broadly divided into three groups in light of current developments in the field of machine learning by leveraging **supervised** and **unsupervised** techniques as well as a novel combination of both. While doing this, I aim to address several biological questions relating to relationship between animals, leadership and association, factors underlying decision making and migration patterns among others. Also from another perspective, a great portion of this thesis is devoted towards studying decision making in animals as it affects movement behaviour across scales.

To be specific, the thesis aims to address several questions defined broadly as:

- Can automatic means of quantifying abnormal movement behaviour in sheep with Batten disease using **unsupervised** machine learning techniques be developed?
- Can recurrent **supervised** deep learning models with backward and forward integration of information be a better model of forecasting migration patterns compared to models without these capabilities?
- Can recurrent **supervised** models with some pseudo objectives be a better model for sequential data

given the presence of an imbalanced distribution of classes?

- Lastly, can a hierarchical recurrent model leveraging **supervised** and **unsupervised** methods be a better model of multimodal movement behavioural data compared to some alternatives?

1.2 Contributions

In this thesis, I have proposed several computational models of animal movement behaviour broadly focused on addressing issues bordering on health and welfare of animals, understanding and modelling the movement behaviour of sheep as well as understanding and forecasting migration patterns and stop over decisions of migratory birds among others. In particular, I made the following contributions:

- In **chapter three**, I showed how simple information theoretic metrics can be used to find patterns in animal movement data across a wide range of context, taxa and scales. This has been demonstrated using publicly available animal movement data from different species across several spatiotemporal scales.
- In **chapter four**, I developed **unsupervised** machine learning and statistical techniques for automatically quantifying abnormal movement behaviour in sheep with Batten disease.
- In addition, in **chapter five**, I developed a bidirectional recurrent deep learning model capable of forecasting fine-grained migration patterns in birds and to be precise Turkey vulture (*Cathartes aura*) using **supervised** methods.
- Furthermore, in **chapter six**, I proposed a bidirectional recurrent network with pseudo loss objectives to forecast stop over-decisions of the Turkey vulture using **supervised** methods.
- Lastly, in **chapter seven**, I developed a predictive and generalisable model of sheep movement behaviour in a flock which has the potential to help understand the decision making process of sheep using both **supervised** and **unsupervised** methods.

1.3 Associated Papers and Abstracts

The contributions in this thesis have either been published, under review or about to be submitted and are listed below:

- **Owoeye, K.A.**, Musolesi, M. and Hailes, S., (2018). Characterizing animal movement patterns across different scales and habitats using information theory (preprint).
- **Kehinde, O.** & Stephen, H. (2018). Online Collective Animal Movement Activity Recognition. Modeling and decision-making in the spatiotemporal domain workshop, NeurIPS 2018, Montreal, Canada.
- **Kehinde, O.** (2019) On the challenges of modelling long term animal behaviour in complex environments. Learning Meaningful Representation of Life Workshop, NeurIPS 2019, Vancouver, Canada. (Abstract)
- **Kehinde, O.** (2020) Forecasting Avian Migration Patterns using a Deep Bidirectional RNN Augmented with an Auxiliary Task. Special track on AI for Computational Sustainability and Human well-being IJCAI 2020.
- **Kehinde, O.** (2020) Preventing Future Outbreaks: A case for global surveillance of wild migratory birds. Workshop on Machine Learning for Global Health, ICML 2020.
- **Kehinde, O.** (2021) Forecasting Stop-over decisions of Migratory Birds using a Deep Bi-RNN with Pseudo loss objectives, IJCNN 2021.
- **Kehinde, O.**, Musolesi, M. and Hailes, S., (2021). Quantifying Unusual Neurological Phenotypes in Collective Phenotypes (preprint).
- Hierarchical Compositional Representations with Relational Memory for Learning to Herd (In preparation).

1.4 Thesis Structure

The rest of the thesis is organized as follows:

- I review related literature in **chapter two** where previous works relevant to the contributions are discussed.
- **Chapter three** is dedicated towards investigating several non-parametric information theoretic metrics towards mining animal movement data, while **chapter four** is focused on the development of automatic

methods for quantifying abnormal movement behaviour in sheep with Batten disease.

- While in **chapter five**, a bidirectional recurrent deep learning approach is proposed towards forecasting migration patterns in Turkey vulture, **chapter six** is dedicated towards forecasting the stop-over decisions of this bird.
- In **chapter seven**, I investigate a novel combination of several deep learning methods towards building a predictive model of sheep movement behaviour in a flock.
- The thesis is concluded in **chapter eight** with further discussions on conclusion and potential future directions vis-a-vis the contributions of this thesis and other related problems.

Chapter 2

Literature Review

In this chapter, I review relevant literature with respect to the contributions in this thesis.

I discuss related literature under three broad themes representative of the major contributions to this thesis highlighted in the previous chapter, in addition to discussions on deep learning methods applied to the study of animal behaviour as well as other literature towards providing some contexts to the understanding of several concepts in this thesis. More specifically, first, I will discuss sensor technology and its role in understanding animal movement behaviour followed by classical techniques for understanding animal behaviour. Then I will discuss related literature in the area of machine learning for animal welfare and anomaly detection, followed by migration and stop-over in migratory birds and then animal behaviour modelling and decision making using artificial intelligence. This chapter ends with a review of concepts relevant to understand some of the methods used in this thesis and their justification followed by deep learning methods relevant to the contributions of this thesis and animal behaviour study in general.

2.1 Sensor Technology & Applications in Understanding Animal Behaviour

A number of sensor technologies such as cameras, GPS, accelerometers, magnetometers and Radio-frequency identification (RFID) have been used in studying animal behaviour. However, since the focus of this thesis

is on animal movement data that have been collected using GPS loggers, I give a brief overview of the technology behind the operation of this sensor and the associated challenges of deploying it to collect animal movement data. Sensors generally operate by converting physical measurements into measurable electrical signals. For instance, the GPS is a sensor technology that uses a network of satellites with microwave signals transmitted to receivers to keep track of location real time using a constellation of at least three satellites. These satellites use their distances to the GPS receiver obtained by transmitting a signal to the receiver, as a radius to construct a circle each with the intersection of all three circles deemed as the the location of the receiver[1]. The first deployment of these devices for animal tracking dates back to the 1990s [5] with more deployments in later years as the manufacturing cost of these devices reduced greatly.

Problems with using these GPS devices for tracking animal movement however include, interference from certain atmospheric conditions including storms and obstacles such as buildings and trees, jamming, battery exhaustion or failure and detachment from the body of the animals when used in the wild. This leads to the degradation in the quality of data collected or in the worst case, the loss of a huge chunk of data. Probabilistic methods that can handle uncertainties can however help alleviate some of these problems and will be explored a lot in this thesis while trying to obtain the necessary utility from these data.

2.1.1 Applications in Understanding Animal Behaviour

Below, I describe the applications of these sensors in the animal behaviour research community under three themes.

2.1.1.1 Understanding Migration Patterns

I review literature where migration patterns of animals have been studied using GPS sensors here. The list though not exhaustive provides an overview of applications in this regard. Just recently, [6] used geolocator tags to track the migration pattern of the Canadian Eastern Whip-poor-will (*Antrostomus vociferous*). They concluded that southeastern united states as well as the Guatemala and Mexico provinces should be considered as potentially important locations for these animals due to their tendency to have stopovers as well as hibernate respectively in these locations. In a similar vein, [7] used geolocator tags to track the migration patterns of red-backed shrikes (*Lanius collurio*) in Western Europe. The outcome of their analysis showed that while these birds fly straight from their non-breeding area in Southern Africa back to Western Europe, they do not fly straight from their breeding ground in Europe to the non-breeding ground during

[1] Also known as trilateration.

autumn. Instead, they fly via south east Europe a strategy alluded to the potential benefit of wind during the autumn although they did not rule out the influence of other factors. Using satellite tags, [8] investigated the influence of wind on the migration patterns of fledgling honey buzzards (*Pernis apivorus*). They showed that 49% of the variability in the bird's longitudinal displacements was accounted for by wind conditions during migration.

2.1.1.2 Monitoring Animal Welfare

Behaviour most especially movement behaviour can be an indicator of the overall well-being of an animal. In this section, I briefly review recent works carried out in this regard where sensors have been deployed to monitor the welfare of animals as this improve accuracy of diagnosis as well as make the work of most farmers less laborious. The weight of pig can be an indicator of their overall welfare and to estimate the weight, [9] used a top-view camera to analyze motion and estimate the weight of pigs without any form of human intervention. Using audio sensors, [10] showed how the sound emitted by a cow can be used to detect oestrus. Similarly, [10] used audio sensors to detect respiratory diseases in a group of pigs. Due to the adverse effect of aggression among pigs on economic returns and welfare, [11] used a kinetic depth sensor to detect aggressive behaviour in a commercial pigpen. Bearing in mind the growing economic cost of stress to the poultry industry due to to its negative effect in the overall wellbeing of chickens, [12] developed a non-invasive audio-visual sensor to notify farmers of a stressful situation in a commercial poultry facility. While in [13] with the use of proximity loggers, the relationship between the social life and productivity of a dynamic group of dairy cows was investigated.

2.1.1.3 Other Applications

Sensors have been deployed in a variety of other applications for studying animal behaviour. For example, [14] using GPS loggers studied how the wandering albatross adjusts its foraging movements at different scales with respect to prey capture. Similarly, [15] with the aid of GPS trackers investigated the role of olfaction in the foraging trips of incubating Scopoli's shearwaters within the Mediterranean sea while [16] studied the effect of density dependent effect on the foraging behaviour of the brown pelican (*Pelecanus occidentalis*) originating from several breeding colonies in the subtropical northern Gulf of Mexico. With the aid of GPS collars, [17] assessed the foraging behaviour of cattle grazing semiarid rangeland in eastern Colorado. With respect to habitat selection and conservation, [18] investigated the influence of vehicular activities on mule deer habitat selection in Western Wyomig using GPS sensors attached to the deer and infrared sensors on

the road. The impact of human development on the movement of wolves was investigated by [19] with the aid of GPS. Migratory route of the wildebeest (Connochaetes taurinus) was mapped out by [20] with the aid of GPS to define areas that are well protected.

2.2 Classical Techniques for Understanding Animal Movement Behaviour

Here, I will discuss existing techniques for studying animal movement behaviour under three themes similar to [21]. These three themes are, quantifying movement patterns, identifying change points and identifying hidden states. I expatiate on this below.

Quantifying behaviour: Quantifying movement patterns of animals is important and useful for measuring a variety of animal movement phenomena. I highlight several important metrics used to quantify animal movement patterns here.

- **Displacement, Step-length [22, 23]:** A measure of the distance covered between two consecutive points.
- **Velocity [24]:** This represents the displacement relative to time.
- **Net-squared Displacement [25]:** Measures the squared displacement between the first and current relocation of a trajectory.
- **Turn-angle [23]:** Quantifies the magnitude of change in direction.
- **Straightness index [26], Tortuosity [27], Sinuosity [27]:** Quantifies how curved an animal's trajectory is.

These metrics and techniques while useful for quantifying a variety of animal movement phenomena, can be inadequate for some other phenomena in light of current deployment of GPS sensors in novel environments for different applications. Therefore, the limitations of some of these metrics will be shown in this thesis and some ways to get around them.

Change points: Identifying behavioural change points is critical towards understanding changes in the behaviour of animals. The use of wavelet was suggested by [28] to identify behaviour segments in animals. The same approach was used by [29] but in their case a discrete version of this wavelet was used to identify

change points in the time series of movement of the Adult Lesser Black-backed gulls as well as [30] who used the discrete wavelet to identify behavioural modes in Turkey vultures (Cathartes aura). A likelihood based method for detecting significant change points was proposed by [31]. This approach applies a sliding window over a continuous time series signal to detect points that are different from the rest. Other method include change point test [32] that detects significant changes in the turn-angle between the starting point and the point of interest in a trajectory.

Hidden State: To reveal the underlying processes behind complex movement data, various state space models have been proposed to identify the hidden states from observational data. The most popular among these model even outside of the animal behaviour research community is the Hidden Markov Model (HMM). For instance [33] showed how to predict behavioural states in the Woodland caribou using metrics such as displacement and turn angle with the aid of the HMM. Due to the variability in the number of hidden states between conspecifics and even animals of different species, [34] used the hierarchical bayesian state space model to analyse the movement data of 14 Leatherback turtles during their southward migration from Canada. To correct for errors in observational data, [35] used a state space model together with Kalman filter to model the internal behavioural state and the proximate movement response of the elk with respect to several environmental variables.

2.3 Machine Learning for Animal Welfare & Anomaly Detection

Machine learning is fast becoming a useful tool for assessing animal welfare and holds promise most especially for automating behaviour measurement for welfare assessment. I discuss related literature in this area here under two themes. First, I will discuss machine learning methods and their applications in animal welfare and then discuss machine learning methods for anomaly detection.

2.3.1 Machine Learning for Animal Welfare

Beyan & Fisher [36] used a hierarchical classifier to detect novel as well as abnormal fish movement in an aquatic habitat. Using features such as velocity, acceleration and turn angle among others, their corresponding principal components were extracted. The principal components were clustered and, together with the labeled data, a hierarchical classier that leverages similarity of data was built and showed good classification accuracy. To detect anomalous patterns in the behaviour of fishes real-time for applications in monitoring the cause and progression of diseases or even death, [37] used a real time object detection deep learning method to

monitor their postures. For aquatic toxicity detection and abnormal pattern analysis in aquatic animals in general using computer vision, see [38] for a synopsis.

Several authors have also proposed different methods to automatically diagnose the health of animals by monitoring their cough using intelligent systems. While Giesert et al. [39] used a Hidden Markov Model to classify coughs into two groups (coughs from healthy pigs and coughs from pigs with respiratory infections), Sa et al. [9], used a motion history image-based method to differentiate the motion peculiar to a coughing animal from other movement patterns. In a similar vein, a neural network was used to analyze patterns in pigs' vocalizations in order to distinguish between sick and healthy animals [40]. While with aid of several machine learning algorithms, an early warning system was developed to detect sick broilers in [41].

In respect of pain assessment and measurement, [42] proposed a deep convolutional recurrent network for pain classification in horses, stressing that spatio-temporal evolution of behaviour is essential in this task. Similarly, to assess post-surgical and post anaesthetic effects on facial expression in mice [43], convolutional neural architectures were proposed using Resnet architectures while in [44], the Inception V3 convolutional neural net was leveraged to predict pain/no-pain in mice using images as input.

More recently, a novel method to detect pain levels in sheep was proposed by [45]. They showed that is is possible to estimate the pain level in sheep using facial features alone. They argued that, by using the Histogram of Oriented Gradients to describe the facial features, followed by training a separate classifier for each feature using a Support Vector Machine, the pain level in Sheep can be estimated. To measure the quality of sleep as well as investigate other forms of neurological dysfunction, [46] carried out an electroencephalography study of sheep with Batten disease. Their results show that sleep abnormalities as well as epileptic waveforms associated with children with Batten disease were also found in these sheep, thus demonstrating the relevance of using these species for *in vivo* studies of human degenerative diseases.

2.3.2 Anomaly Detection

Here, I review methods on machine learning methods for anomaly detection[2] looking at modern and old approaches. While old methods are based on classical machine learning methods, modern ones leverage recent advances in deep learning. For extensive review of deep learning methods for anomaly detection see [47]. Anomaly detection methods can be grouped into supervised[3] and unsupervised models[4]. However, only the

[2]Anomaly detection can also be referred to as outlier/novelty detection.
[3]Supervised methods use previous instances of anomalous and non-anomalous data to train a model usually a classifier.
[4]Unsupervised models leverage the measure of surprise in a data with respect to a population to determine anomalous data points.

latter will be discussed here as they are most relevant to the methods I aim to use in this thesis. I will also discuss deep learning methods for anomaly detection.

2.3.2.1 Unsupervised Anomaly Detection Methods

According to [48], unsupervised anomaly detection methods can be classified into five groups below.

- **Clustering Methods:** These set of methods [49, 50] are used to find similarities in data by grouping similar datasets together. One key assumption using these methods is that anomalous data instances belong to the small clusters while the normal data points belong to the bigger clusters. In this thesis, these methods will be considered and investigated.
- **Nearest-neighbour Methods:** These methods [51, 52] uses a similarity metric such as the euclidean distance to compute the distance between all data points with the assumption that normal data points are closer in space than abnormal ones. One problem with using these methods is the choice of the number of nearest neighbours to be considered.
- **Statistical Methods:** It is assumed using these class of methods [53, 54] that a data point is anomalous if it's not generated by an assumed model. The problem with using these methods in the real world is that, in some cases, this assumption doesn't hold true perfectly. In this thesis, some of these methods will also be considered and investigated.
- **Information Theoretic Methods:** The information theoretic methods [55–57] for anomaly detection uses concepts of entropy and complexity to look out for patterns in the information content of a data. However, it can be difficult to detect anomalies in the data unless they are present in fairly large quantities.
- **Spectral Methods:** The key idea behind these methods [58, 59] is that, the data is embedded in a lower dimensional space where the anomalous data points can be found easily. The problem with using these approaches naively is that the normal and abnormal instances may not be separated easily sometimes as expected in the lower dimensional space.

For comprehensive information and description of the aforementioned methods, see [48].

2.3.2.2 Anomaly Detection with Deep Learning

While deep learning models will not be used for anomaly detection in this thesis due to the quantity of data available, I will discuss briefly relevant literature in this area. Generative Adversarial Networks (GANs) [60] and Auto encoders [61] are two of the most popular deep generative models[5]. I will discuss the application of both for anomaly detection.

GAN was used for for instance in [62] to identify anomalous image data using a residual loss function coupled with the GAN loss objective to compute an anomaly score which expresses a fit between normal generated images and a new query image. The auto-encoder on the other hand, leverages the reconstruction error to identify anomalous data samples where higher reconstruction errors are expected for abnormal samples and lower errors for the normal ones. See [63, 64] for instance for examples.

2.4 Forecasting Animal Migration and Stop-over Decisions of Migratory Birds

Forecasting animal migration and stop-over sites/decisions have some history but has received much attention recently. Due to the proliferation of low cost sensors and advances in telemetry, useful datasets of animal migration are becoming increasingly publicly available to enable animal behaviour researchers use a data driven approach towards answering important biological questions regarding animal migration. In this section, I discuss previous attempts in this area. In addition, I will also discuss relevant works in the role of wild birds in spreading infectious diseases.

To mitigate the mortality of aquatic animals, to be specific European eels (*Anguilla anguilla*) due to harm from hydroelectric power plants, [65] proposed a seasonal autoregressive integrated moving average to forecast their migration in advance using datasets of silver eel migration from two fishing sites in North-western France. The idea in this work was to minimize mortality by shutting down turbines. However, to minimize down time as much as possible and, by extension, losses to hydropower producers, they argued it is imperative to be able to precisely forecast the peak migration times of these animals.

Recently, [66] built a predictive model of spring migration intensity of birds in North America using radar and weather data with the aid of XGboost. In a similar vein, [67] used radar measurements of bird densities in

[5]Generative models are models that tries to learn the underling distribution of a variable.

the Netherlands to predict their migration intensities for applications in improved flight safety. The problem with predicting migration intensities however is that, it cannot be guaranteed to be optimal for managing some of the critical infrastructures causing these deaths as different birds potentially have different migration timing across seasons.

Even more recently, evidence for a phenological shift due to climate change was provided by [68] using 24 years of archived radar data in North America. Their results revealed advancement in peak spring migration by 0.6 days per decade with the effect highly correlated with latitude. While it is unclear from their analysis where and when some of these changes occurred, as well as how it varies across species, tracking bird movement at a finer scale as per the data used in this work has the potential to shed more light on some of these issues. Similarly, the relationships between short term weather conditions to daily migration intensities of different bird species were examined by [69]. Their analysis revealed how different populations and species respond to weather conditions in varying ways while a maximum number migrate when the average weather condition is fair.

Previously, using the dataset of white-fronted Geese collected over multiple years, [70] attempted to forecast the stop-over sites of these birds with the aid of a Markov chain, using historical information. There are, however, problems with this approach. First, forecasting stop-over sites that have never been used before becomes difficult and, second, estimating the duration of stop-over is impossible, which can be useful for planning purposes for instance by health authorities for the purpose of preparedness and response planning. In a related work, [71] used random forest to determine important variables such as roads, ecotone, agricultural land, among others, relevant to stop-over in the migratory whooping crane (*Grus americana*) for conservation purpose while [72] reviewed various largely qualitative methods for estimating stop over duration in migratory birds.

The role of migratory birds in spreading pathogens cannot be overemphasized. Recently, [73] showed the avian influenza virus (AIV) migration rates are higher within flyways than between using a large scale data of gene segments from multiple geographical locations. In a similar vein, [74] used information on phylogenetic associations of virus isolates, migratory bird movements as well as trade in poultry and wild birds to estimate the spread of the highly pathogenic avian influenza in several locations across the world. Also, [75] detected the AIV in wild ducks moving between Europe and Asia using swab samples and fresh droppings. While the approaches described here are mainly in vivo and partially in silico, the one proposed here is mainly in silico. Lastly, [76] used serological data with environmental features to model the spread of the West-Nile virus in

Greece using Geographical Information System technology .

2.5 Models of Behaviour and Decision Making in Animals

Recent advances in computational intelligence, most especially in machine learning, have accelerated research in the modelling of animal behaviour for the purpose of understanding their decision making processes with the development of tools that have made it easy for example to automate the tracking of animals over long recordings and to annotate behaviour over these recordings [77]. The majority of the methods relevant to this field can be classified into supervised and unsupervised methods. While unsupervised methods allow the classification of behaviours without any prior knowledge regarding these behaviours and can be used to detect behaviours that are novel, supervised methods allow the classification of behaviours based on prior knowledge using some similarity metric [77]. There is a range of work however in this area and I discuss some of it here.

In their bid to understand the neurological basis of mouse behaviour, [78] used 3D imaging with an auto regressive Hidden Markov Model to learn the strategies these animals use to adapt to their environment. Their results suggest that mouse behaviour is organized into sub-modules combined in such a way as to create a coherent sequence of actions. They further showed that there may be three strategies[6] that may be used by the brain in implementing behavioural adaptation.

Recently, the authors of [79] tried to understand the decision making process of *C. elegans* from time series data with the aid of inverse reinforcement learning to learn the reward function. Using a Linearly solvable MDP [80], they were able to identify two main strategies used by fed worms, namely directed migration to reach a specific temperature and isothermal migration to move along a constant temperature. Starved worms, in contrast, use only the absolute temperature for moving away from temperature. Similarly, with the aid of deep reinforcement learning, the authors of [81] used a virtual rodent to study motor activities across several contexts while the rodent solved several cognitive tasks. Their experiments revealed the model encoded task strategies using two representations. The problems with these body of work is that, the environment used was either simulated or laboratory based and does not reflect the behaviour of these animals in the real world.

With the aid of deep generative models, the authors in [82] were able to model the sensory-motor relationship that exists between fruit flies (*Drosophila melanogaster*) and their environment using a method

[6]Transition between behavioural modes is non-stationary, new behaviour can be created by changing the transition dynamics between modes and there are adaptations in space as to where certain behavioural modules are activated

that simultaneously classifies action whilst predicting motion. Still working with fruit flies, [83] attempted to model the sensory processing of stimuli with respect to behaviour. The limitations of these approaches are similar to the earlier ones discussed in terms of the challenges in generalising from restricted and artificial environments in that, the state space is limited and the actions are relatively uniformly distributed. In environments in which the animals being modelled exhibit different movement patterns that are skewed, live in co-operative or competitive societies, these methods will perform poorly.

Finally, the authors of [84] modelled the behaviour of dogs using visual information collected from cameras as input and inertial measurements of the limbs, tails and the back of the dog as output. The model, a recurrent network with the support of pre-trained convolutional architectures directly predicts the actions of the dog from clusters of the inertial measurements. Their aim was to learn visual intelligence using a combination of behavioural data from these sensors.

2.6 Model & Concepts: Background and Review

Here, I will discuss relevant concepts and terminologies needed in understanding some of the methods used in this work most especially as it affects the work on building models of decision making in animals and the motivations behind them.

2.6.1 Interpretability

Understanding the decision making processes of modern deep learning systems has been of importance in recent times and is necessary to enable these systems to be deployed in safety critical systems. While these systems use a connectionist approach towards learning, it is often unclear how they arrive at their decisions. Symbolic reasoning which involves specifying hard coded rules has been proposed as alternatives but it has its own disadvantages too as things can become easily complicated given very difficult problems. One way forward is to find a balance in the usage of both approaches by specifying soft and not hard coded rules or priors that can be learned by neural networks[7] and connecting them using architectures such as hierarchical ones that allow these systems to have efficient representations vis-a-vis their decision making ability. I discuss more on the concept of hierarchical representations in the section below.

[7]Also known as neuro-symbolic AI.

2.6.2 Hierarchical Models

These are statistical models that exist at different levels of hierarchy with different parameters. While there is evidence that suggests that human behaviour is hierarchical [85], recent evidence also suggests the mouse behaviour may be hierarchical [78] and perhaps other animals as well. Due to the complex nature of the behaviour of animals in their natural environment, these behaviours are better modelled using a hierarchical framework where behaviour may be perceived at different levels of resolution and scale with respect to time, activities and actions. This hierarchical modelling aside its role in cognition [86], offer many advantages, enabling the learning of multiple tasks and dependencies [87] and as a consequence is able to prevent problems like catastrophic forgetting [88–91] associated with modern neural networks while learning sequentially multiple tasks.

2.6.2.1 Optimal Behavioural Hierarchy

While behaviour have been shown to have a hierarchical structure, recent work [92] has shown that all hierarchies are not created equal. The optimal hierarchy as such is one that encourages behaviour to be represented simply, ensuring quick adaptation in novel states or environments, also known as meta learning [93, 94], while also enhancing learning and planning. In this dissertation, first, I will show why architectures might be important for meta learning and also investigate and show why different hierarchical structures provide varying representations with respect to learning tasks.

2.6.3 Innate behaviour

It has been suggested that animals and humans have structured brain connectivity and wiring [95, 96] that is encoded in the genome and is passed from one generation to another. It is this wiring that perhaps enables rapid learning and allows some animals to possess some intelligent behaviours shortly after birth. The objective of this thesis with respect to decision making is therefore akin to learning some of the structures of this internal wiring as it affects movement behaviour in sheep.

2.6.4 Modularity, Compositionality and Relational Reasoning

Solving complex and long range problem or tasks requires the development of skills or modules (modularity) [97, 98] that are reusable. The notion of compositionality implies complex systems are composed of simple modules that are reusable over time [99] and the capability of finding relationships between modules

(relational reasoning [100]) is central to advanced intelligent behaviour. In this thesis, with respect to decision making in animals, ideas from these three concepts are going to be investigated where the original movement behaviour is decomposed into smaller behavioural modules using tree prior/structures where each child node is connected to its immediate parent node via supervised learning (relational reasoning) and the whole network uses these connections holistically to determine what modules to use (compositionality) given any stimuli.

2.6.5 Chunking/Clustering

To construct the modules described above, clustering also known as chunking is necessary. In a famous and widely cited paper [101], chunking has been argued to be important in the memory for organization and processing of information/sensory inputs. Meaningful chunks/modules as posited in this paper are useful and I will show in this thesis why this is true.

2.6.6 Working and Short Term Memory

To see connections in time series data, working memory is important [102]. Working memory is short term memory with limited capacity that enables the storage of information to guide decision making and behaviour [102, 103]. In this work, I will use recurrent neural networks with a small number of cells to connect parent nodes to the child nodes (factorized behaviour) so as to guide received stimuli to an appropriate behavioural module containing the relevant action to take[8].

2.6.7 Activity Recognition, Event Segmentation & Perception

Segmenting continuous behaviour or activities into meaningful events has been shown to be an important core component of perception with implications in the neural processes underlying memory and learning [104]. Modern deep learning systems using completely unsupervised methods in form of auto encoders to learn complex tasks struggle with learning semantically meaningful representations most especially over long temporal horizons. For instance, with respect to human behaviour modelling, the sitting posture used at home by an individual might be assumed to come from the same latent space as the sitting posture used in the workplace. Segmenting behaviour along temporal dimensions with respect to a relevant meaningful context is therefore useful to avoid problems of this nature. In this work, to ensure meaningful flow of information coming from sensory inputs, the concept of collective animal movement/event recognition will be proposed

[8]These recurrent neural networks serve as relational memory networks tracking the relationships between entities in the hierarchy.

and used as a prior. The goal here will be to recognize and classify the movement behaviour of the flock in such a way as to make learning and representation efficient and meaningful.

2.6.8 Imitation learning & Behaviour cloning

Imitation learning and behaviour cloning methods aim to copy behaviour in a given task. An agent here is trained to perform a task by learning a mapping from observations to actions from demonstrations. This is a supervised learning method and, in contrast to reinforcement learning, does not need the specification of a reward function. The problem with this approach is that state domain is limited, and when an agent finds itself in states not experienced during training (out of distribution problem), it gets lost. This problem is partially solved via reinforcement learning where an agent is trained to explore even though it is expensive. Recent practices, however, have leveraged the strengths of the two methods to obtain superior results [105–107]. The approach to be proposed in this work aims to build on some of the ideas in these methods in combination with new ones to build models that can generate behaviour in more complex environments over longer horizons.

2.6.9 Models

I give a brief overview of all the models that will be used here for clarity. These models belong to two broad categories of supervised and unsupervised learning methods or, from another perspective, discriminative and generative models.

2.6.9.1 Gaussian Mixture Model (GMM)

This is an unsupervised learning method that disentangles data into several clusters using a soft clustering approach with a Gaussian prior using the expectation maximization algorithm. The output is a function of several Gaussians each with a mean μ_k that defines its centre, a covariance matrix Σ_k that defines the width and a categorical probability π_k that defines the likelihood of a cluster where $\sum_{k=1}^{K} \pi_k = 1$. In this work, the GMM will be used to cluster the movement behaviour for easy processing and representation by the mixture density network (see below for description of this network) since the output of this network is parameterised by a mixture of Gaussians.

2.6.9.2 Feed Forward Networks

Also known as multilayer perceptrons, these class of neural networks map a fixed size input to a fixed size output without any form of feedback in between by computing a weighted sum of the input while moving it

from one layer to another via several layers of non-linearities defined by the activation functions [108]. Here, it is the case that there is one input, one output and at least one hidden layer. When there is a feedback in between the output and the input, the resulting network is called a recurrent neural network and I discuss that below. In this work, the feed forward network is used to process the inputs used to learn the parameters of the output of the mixture density network described below.

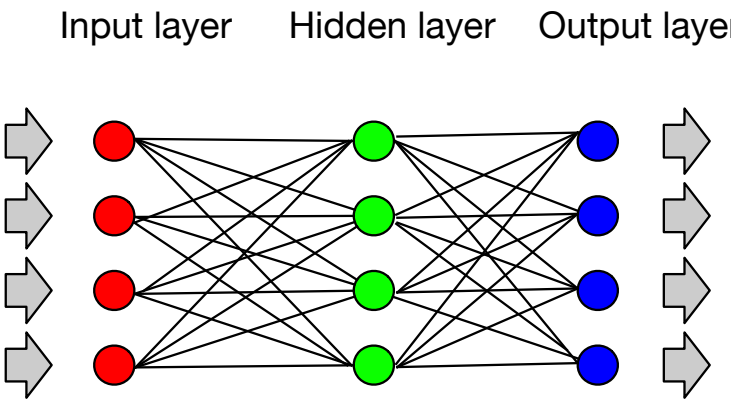


Figure 2.1: **Feedforward network with the input, hidden and output layers.**

2.6.9.3 Recurrent neural networks (RNNs)

Deep RNNs have been shown to be useful in classifying time series and have been successful in handwriting recognition [109] as well as speech recognition [110]. They are a rich class of models with the ability to model long sequences. The Long Short-term Memory (LSTM)[111] network is a variant of the RNN designed to store and access information better than the ordinary RNN. This variant will be used in this work. The LSTM recurrent network is defined by the equations below:

$$\begin{aligned} i_t &= \sigma(W_{x_i}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i) \\ f_t &= \sigma(W_{x_f}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f) \\ C_t &= f_tC_{t-1} + i_t tanh(W_{x_c}x_t + W_{hc}h_{t-1} + b_c) \\ o_t &= \sigma(W_{x_o}x_t + W_{h_o}h_{t-1} + W_{co}c_t + b_o) \\ h_t &= o_t tanh(C_t) \end{aligned} \tag{2.1}$$

where i_t, f_t and o_t represent the input, forget, and output gate respectively. $W's$ are the weight matrices, x_t is the current input data, h_{t-1} is previous hidden output, C_t is the cell state and σ is the logistic sigmoid

function while $b's$ represents the biases in the network.

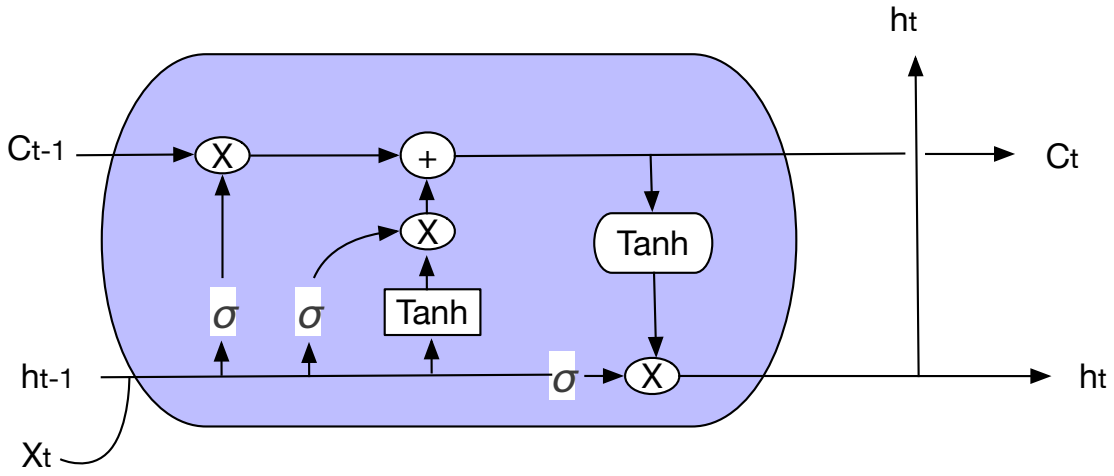


Figure 2.2: **The long short term recurrent neural network.**

2.6.9.4 Discriminative Models & Generative Models

Discriminative models are a class of supervised models that learn to discriminate between different kinds of data. Generative models, in contrast to their discriminative counterparts capture the underlying data generating process. Given two distributions X with labels Y, the discriminative model describes the conditional probability $P(Y|X)$ while generative models captures the joint probability distribution $P(X, Y)$. Some of the models described above can either be deployed in a discriminative or generative framework with both to be considered in this thesis.

2.6.9.5 Mixture Density Network (MDN)

The MDN proposed in [112], is a supervised generative model that leverages the representational power of neural networks to parameterise a mixture distribution. Here, it is assumed the data is generated from a mixture of Gaussians. A portion of the neural network output is used to learn the mixture weights while the remainder is used to learn other parameters such as mean and standard deviation for each mixture component. The mixture weight outputs are normalised with a softmax function so they form a valid discrete distribution that sums to one. The mixture density network can be described using the equation below where the parameters $\pi_{\mathrm{i}}(x)$, $\mu_i(x)$ and $\sigma_i(x)$ represents respectively the mixture weights, mean and standard deviation of the i^{th} Gaussian component and will be determined by the neural network. Y here represents the actions, in this case a bivariate Gaussian of the velocities of sheep movement conditioned on the states X.

$$P(Y = y|X = x) = \sum_{i=1}^{k} \Pi_i(x)\mathcal{N}(y, \mu_i(x), \sigma_i(x)), \sum_{i=1}^{k} \pi_{\mathtt{i}\ (\mathtt{x})} = 1 \tag{2.2}$$

2.7 Deep Learning Methods for Studying Animal Behaviour

While deep learning has revolutionised a lot of scientific endeavour, its application in the area of animal movement is limited. Here, I review relevant literature in the area of deep learning methods for applications in the field of animal behaviour inclusive of animal movement behaviour[9] under two broad themes of feed forward and recurrent neural networks, as well as convolutional neural networks (CNNs)[10]. For comprehensive review of deep learning methods, see [108].

2.7.1 Feed Forward and Recurrent Neural Networks

Recently, deep learning models including recurrent and feed forward network models were used to simulate the movement behaviour of Ant and the migratory black-backed gulls (*Larus fuscus*) [113]. Their analysis revealed the Long Short Term Memory (LSTM) recurrent network performed best for long term simulations amongst the deep learning models.

To predict diving in seabirds and by extension prevent global declines in marine ecosystems [114], a feed forward network was used to model the diving behaviour of these birds with the problem formulated as a classification task. In another work [115], deep feed forward networks were used to predict the overall dynamic body acceleration using bio-logging tags attached to the body of the white shark. Motionless, walking and trotting movements as well as gait behaviours were classified with the aid of feed forward neural networks and wireless sensor networks for the study and remote monitoring of wildlife [116].

In this thesis, I will show the limitations of LSTM in modelling varied and biased animal movement behaviour over longer horizons, and show how to overcome these limitations whilst investigating the performance using several quantitative metrics over multiple datasets and context with respect to several animals.

[9] I consider literature that have never been mentioned previously in this review.

[10] Deep learning models can be classified majorly into feed forward networks, recurrent neural networks and convolutional neural networks.

2.7.2 Convolutional Neural Networks

To understand the functional relationship between brain and behaviour in animals, [117] investigated the performance of three convolutional neural network architectures to model the behaviour of mice performing several tasks. Most importantly, they showed the importance of transfer learning using pre-trained models in accelerating the training of the convolutional neural networks for behaviour modelling. For pose estimation in animals with applications in connecting neural activities with behaviour, CNN was also used on videos of freely behaving fruit flies and mice to track points representative of pose of the head, body, wings and legs [118].

In a bid to detect the presence or absence of whales in water bodies, [119] used satellite imagery with convolutional neural networks leveraging pre-trained ResNet model. The task of identifying and re-identifying bird species in both captive and wild contexts using CNN was considered in [120], using images contextually different between the train and test data. In a similar vein, CNNs were used in [121] to identify individual brown bears (*Ursus arctos*) for applications in conservation. Deep convolutional networks were also used in [122] to classify several bird species using image patches located by the pose with their results showing that a model that integrates lower-level feature layers and higher-level feature layers with unaligned image features with pose normalization works best.

CNN has also been used to recognize and classify the behaviour of Drosophilas to know whether they were on egg-laying substrates or not in contact with the substrate using frames from videos [123]. The focus of this thesis however is to classify movement behaviour using mostly recurrent neural networks. But, I aim to use a pre-trained convolutional neural network (CNN) model to map the stop-over habitats of some migratory birds so as to understand the factors influencing their choice of stop-over sites.

2.8 Conclusion

In this chapter, I have reviewed previous literature under all the themes relevant to the contributions of this thesis. This thesis aims to address some of the problems with existing work or tackle new ones that have not been considered previously. More specifically:

- With respect to the welfare of animals, all the literature discussed above differ from the one of interest in this thesis in that, first, I am interested in finding abnormal movement behaviour in sheep using data

collected with non-invasive means. Second, I am using fully unsupervised machine learning methods compared to previous works that have used mainly supervised methods.

- Regarding migration, while previous literature have largely focused on forecasting migration intensities across a huge taxa, in most cases birds, analysing long-term migration patterns at a more fine grained scale will be key towards shedding more light on some of the problems associated with using coarse-grained data as obtainable in most of the literaure. Also, I will consider multiple fine-grained migration states[11] across two continents and formulate my prediction problem as a classification task which is more optimal for managing critical infrastructures when used with more fine grained migration data. I will also investigate in this thesis, new environmental and weather variables to know which are most influential with respect to the migration states, using mutual information.

- With reference to models of stop-over in migratory animals, the work in this thesis differs in that while some of the methods proposed are majorly qualitative, the work in this thesis is purely quantitative. In addition, the task of estimating the stop-over length in this thesis will be considered from a forecasting perspective compared to some of the reviewed literature that formulated the problem mainly as a one-to-one classification task. Also, pre-trained convolutional neural network will be used towards understanding the nature of the stop-over habitat using the land cover paradigm.

- And lastly, concerning models of decision making in animals, most of the works that have been carried out in this field have been restricted to well controlled laboratory environments where the natural behaviour of animals are highly restricted and the behavioural dynamics simplified. As recently argued by [124], the use of unrestrained natural behaviour of animals is essential in understanding the true neural basis of their behaviour. The work in this thesis will attempt to do just that by modelling animal behaviour in more natural settings using modern machine learning methods. In addition, while previous literature have focused on using a flat clustering to develop movement primitives, I aim to use hierarchical clustering with a tree prior in this work and a relational memory to connect levels together so the model can generalize over long horizons outside of laboratory environments. Also, evaluation will be carried out over longer horizons compared to the short horizons used in previous works.

In the next chapter, I will review and investigate the application of several information theoretic metrics towards mining animal movement data.

[11] Compared to just one migration state considered in previous literature.

Chapter 3

Characterizing Animal Movement Patterns Using Information Theory

Here, I review and discuss potential applications of key information theoretic metrics towards understanding and finding relationships in animal movement data considering most of the contributions of this thesis involve either understanding the way animals process the information arising from the sensory inputs they receive or, how they respond to this information.

3.1 Introduction

Information theory has always played an important role in biology [125, 126]. It is a field that is devoted to studying the storage, communication and quantification of information founded by Claude E. Shannon in his influential paper [127] and lies at the interface of mathematics, statistics, computer science and electrical engineering. While initial research in this field was mainly theoretical, a plethora of practical applications have been witnessed in the past decades. For example, concepts and techniques from this field have been used in several fields such as neurobiology [128], pattern recognition [129], cryptology [130], bioinformatics [131], quantum computing [132] and complex systems [133, 134] with significant success.

Recently, due to technological advances, low cost miniaturized sensors have been increasingly adopted for tracking the behaviour of animals across different scales and habitats. These sensors include, but are not

limited to, Global Positioning System (GPS) receivers, accelerometers and radio-frequency identification (RFID) tags. This has led to an explosion in the deployment of these sensors in different habitats, across scales by animal behaviour researchers. A natural consequence of this development is that it is now possible to try to quantify and understand a variety of aspects related to animal behaviour such as migration patterns and routes, feeding, reproduction and mating patterns, conservation, monitoring of endangered species, epidemic spreading, resource use, social behaviour and association. In the field of animal behaviour, GPS, accelerometers and cameras are the predominant sensors deployed to measure several behavioural properties of animals. I limit the discussion here to GPS sensors due to their ease of deployment and the apparent simplicity of interpreting the data they produce as well as their consequent popularity relative to other sensors. For example, GPS sensors have been used to study selfish herd behaviour of sheep under threat [135], the hierarchical structures of group dynamics in flocks of pigeons [136], migration patterns in vultures [137], productivity in cows [138], and social relationships in birds [139] just to name a few.

Information-theoretic approaches can provide *complementary* insights in the study of animal movement behaviour. In other words, these approaches do not replace the existing ones, but they are able to provide additional information about animal behaviour patterns, especially in terms of movement, which are not apparent using other types of analysis. Probabilistic approaches are also usually more robust in presence of noise, a common feature of sensor data. The number of applications of concepts and techniques from information theory to analysis of animal movement in the literature is limited. However, information-theoretic metrics have been used in the past for example to study predator-prey relationships [140, 141] in animals.

Here, I discuss how three classic information theoretic metrics, namely Shannon entropy [142], mutual information [143] and Kolmogorov complexity (normalized compression distance) [144] can be effectively applied to the study of animal movement and provide additional *complementary* insights about animal behaviour [55]. In other words, I explore how they can be used as tools for studying animal movement data. Indeed, the goal of this work is not to introduce new metrics, but to demonstrate the potential in using information theoretic concepts to understand animal behaviour. I introduce each metric separately and then I discuss how each metric can be applied to a practical problem, by discussing a case study in detail. More specifically, I demonstrate how these metrics can be used to characterize the movement patterns of animals across different scales and habitats. It is worth noting that these methods do not provide new ground-truth information, but they allow for identifying emergent patterns and formulating hypotheses that can be verified for example by means of further experimental observations in the field.

3.2 Shannon Entropy

3.2.1 Overview

The information content of a random variable is defined by the Shannon entropy [142] as a measure that quantifies the level of uncertainty embedded in such variable. In the case of animal movement, Shannon entropy provides a useful quantification of the level of regularity and predictability of the movement of an animal.

More formally, it can be defined as the uncertainty associated to a random variable X with realization x, which can be described by the following equation:

$$H(x) = -\sum_{x} p(x) \ \log \ p(x). \tag{3.1}$$

where $p(x)$ is the probability density function and the summation is taken over all possible realizations of X. The base of the logarithm is not important and can take any value, provided the same base is used throughout the analysis.

Most of the information theoretic measures that exist today, some of which I will discuss below, are derivations of the Shannon entropy. However, there are other several definitions of entropy such as Rènyi entropy [145] and Tsallis entropy [146]. The discussion of these metrics is outside the scope of the present work.

Due to the noisy nature of most datasets, probabilistic metrics are becoming increasingly useful for modelling not only animal movement data but in general real world datasets to account for any form of uncertainty inherent in datasets of this nature such as missing data. Therefore, entropy can be used for assessing the overall welfare and well-being of animals instead of a metric like distance travelled. Most animals are known to have a regular activity-rest pattern except under highly unfavourable conditions or when they have some sort of impairment in their general well-being. This implies that animals are supposed to have a relatively high entropy except for the period of harsh conditions when they are either aestivating or hibernating. For this reason, entropy can be used to characterize the movement patterns of animals so as to assess the state of their health. In addition, it can also be used in lieu of tortuosity to describe how tortuous an animal's path is using the turn angle as the input. The conditional entropy is an important element of the

conditional mutual information and can be used for example in understanding swarm behaviour [147].

In the following subsection, I will consider a case study illustrating a possible application of Shannon entropy to the study of animal behaviour and, more specifically, to the characterization of the movement patterns of sheep with neurodegenerative diseases.

3.2.2 Case study: Shannon entropy as a tool for characterizing movement patterns of sheep with neurodegenerative disease

The detection of abnormal locomotion patterns is essential for the early diagnosis of a number of neurodegenerative diseases such as Batten disease in animals. Sheep with neurodegenerative diseases such as Batten disease are known to exhibit repetitive behaviours [46] over time due to gradual loss of motor skills [148] and social awareness [149]. Here, I used Shannon entropy to characterize the movement patterns of a flock of sheep comprising sheep with a natural mutation for Batten disease and their age matched control (mean age of 2 years) group using the dataset of [150]. I used the trajectory of each sheep sampled every second and computed the distance covered every ten minutes over eleven hours (20:00-07:00) each day for a total period of six days. This time window was chosen in order to minimize the influence of external environmental noise in the dataset. I further binned the resulting distance calculated in order to assign to each sub-intervals symbols. To bin the data, given that the data has a skewed distribution, I used the head/tail classification rule introduced in [151] resulting in 12 bins. First, I evaluated this approach on synthetic data where I created two synthetic sheep sampling their distance covered from some random generalized pareto distributions. I created ten of these distributions with different parameters. The abnormal sheep was allowed to sample from just two of these distributions while the normal from all. Details of this can be found in Appendix A. The entropy was calculated and results (Figure 3.1A) show the normal sheep had a higher entropy compared to the abnormal sheep. Furthermore, the entropy for each sheep was computed (see supplementary material below) as well as the mean entropy (Figure 3.1B) for the two groups of sheep. Results show that the Batten sheep on the average had a lower entropy than the normal sheep with p-values (ANOVA) (0.0076, 0.1042, 0.2628, 0.0065, 0.0234, 0.0205) across the six days respectively. The potential impact of uncontrollable environmental variables such as unfavourable weather conditions is significant and may influence the behaviour of the sheep especially because the experiment was carried out in an open field. Therefore, the result should be interpreted with caution. I compared the entropy of the two groups of sheep with their respective average distance covered in (Figure 3.1C) and its mean variance (Figure 3.1D). The Batten sheep can be seen to have

covered, on average, a longer distance over the period of observation thus giving a result that appears to be uninterpretable compared to entropy.

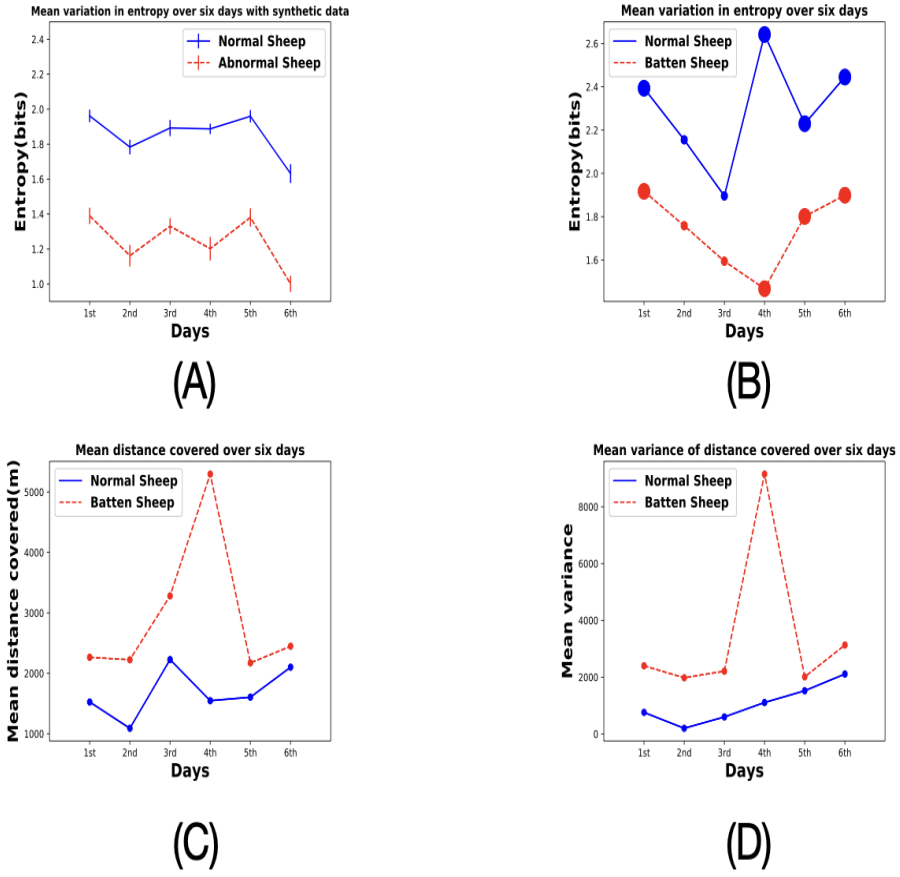


Figure 3.1: **Entropy, distance and variance comparison** (A) Mean entropy with synthetic data; (B) Mean entropy of the two groups of sheep across 6 days where the larger circles represent days when the mean difference in entropy are statistically significant (One way ANOVA with $p \leq 0.05$). The sheep affected by Batten disease can be seen to have a lower entropy due to the tendency to repeat the same behaviour over a long period of time; (C) Mean distance covered by the two groups of sheep; (D) Corresponding variance of the mean distance covered.

3.3 Mutual Information

3.3.1 Overview

I now describe another classic information theoretic measure intimately linked to entropy, called mutual information [143]. The mutual information of two random variables X and Y defines the mutual influence one variable has over the other. Specifically, it quantifies the amount of information in one variable embedded in the other. For this reason, mutual information can be used as a measure of association or social-grouping, for example in the characterization of leader-follower relationship, group coordination and, more generally, collective behaviour [152]. This can be further used to derive a social network [153] and its complementary to the gambit of the group approach [154]. Mutual information can be used to measure non-linear relationships between two variables. More formally, the mutual information of two discrete random variables X and Y, with realizations x and y respectively, is given by:

$$I(X;Y) = \sum_{y} \sum_{x} p(x,y) \ \log \ \frac{p(x,y)}{p(x)p(y)}. \tag{3.2}$$

In this, $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively and $p(x,y)$ is the joint probability distribution function of X and Y.

In the case of continuous random variables:

$$I(X;Y) = \int_{Y} \int_{X} p(x,y) \ \log \ \frac{p(x,y)}{p(x)p(y)}. \tag{3.3}$$

I now consider a potential application of mutual information, in its application to the study of association in pigeons.

3.3.2 Case study: Mutual Information for measuring association and leadership in pigeons

As mentioned earlier, the de-facto method used in the animal behaviour community for measuring association is the gambit of the group otherwise known as co-location [154]. To detect significant associations and minimize co-location by chance, a permutation test (random shuffling of associations) is often carried out

[155]. However, it is difficult to identify a general method for defining co-location. Also, the directional correlation delay time method[1] used by [136] in reconstructing pigeon flight network structure can only detect linear relationships leaving the non-linear relationships undetected.

I demonstrate how Mutual Information can be used to overcome the limitations associated with the methods above by using it to measure association between pigeons in flight. I used the dataset of [156, 157] and select flight 8 as the result was discussed in the literature in detail. I used the time-series of the turn angle of each bird followed by the pairwise mutual information of these time series of the nine birds involved in the flight to obtain a distance matrix (see supplementary material below). As expected, there will always be a certain degree of association between all the birds in the flight, I used a randomization test to determine a threshold for significant pairwise mutual information (see Appendix A). I further build a social network (Figure 3.2A) to visualize the flight formation and the results appear consistent with two previous studies on pigeon flight. First, I observed that pigeons do show a hierarchical formation when in flight as seen in (Figure 3.2A). This is a result consistent with the observations in [136]. Also, I was also able to detect the leader as node M during the flight which is the node M with one edge (Figure 3.2A)[2] (see video of ground truth trajectory). This result is also consistent with the ground truth in the literature [156]. I compared this approach with three other methods: correlation coefficient, transfer entropy and Granger causality (Figure 3.2D-G). While it is not straightforward to compare the performance of the four methods, one basis for comparison concerns leadership. Mutual information shows the best performance in terms of leader identification accuracy followed by the correlation coefficient. I attribute the poor performance of the transfer entropy and Granger causality to the continuous change in positions of the birds when flying.

3.4 Kolmogorov Complexity

3.4.1 Overview

Finally, I describe a similarity metric influenced by Kolmogorov complexity (KC), a metric with foundations in the field of algorithmic information theory. The Kolmogorov complexity of an object represents the shortest computer program that produces the object as output [144]. More formally, the KC of a string x with respect to a reference machine U is defined as:

[1] Measures how long it takes for one bird to change direction relative to another
[2] https://www.youtube.com/watch?v=19srQkTVTGE

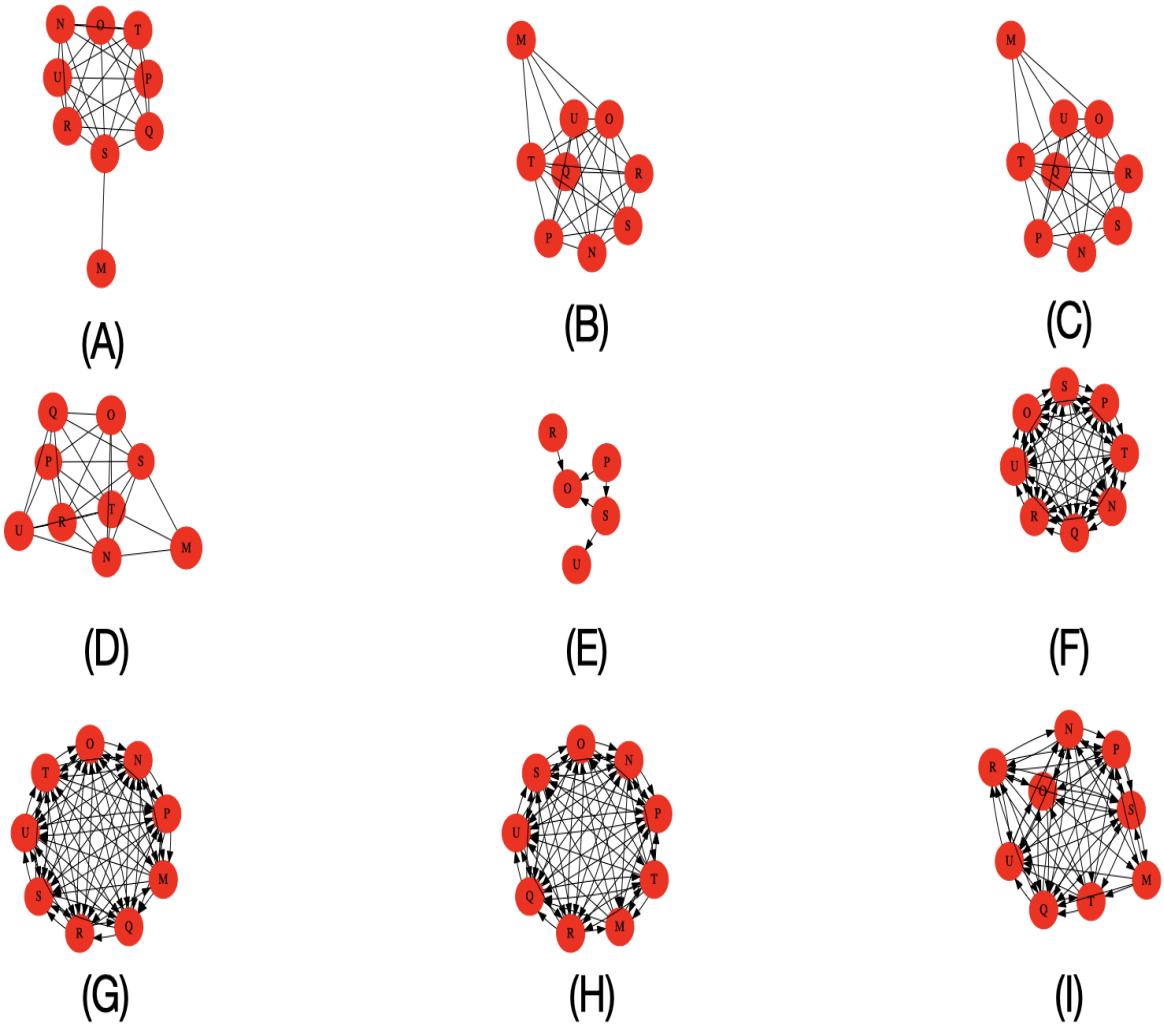


Figure 3.2: **Flight dynamics of homing pigeons.** Characterized by (A) Mutual Information(Kernel Estimator)[158] (B) Mutual Information(Discrete & Continuous mixtures) [159] (C) Mutual Information(Kraskov Estimator)[160, 161] (D) Correlation Coefficient (E) Transfer entropy (Kernel estimator) [161] (F) Transfer entropy (Kraskov estimator) [161] (G) Granger Causality (p-value = 0.05) (H) Granger Causality (p-value = 0.001) (I) Granger Causality (p-value = 0.0001). The tables presenting the pair-wise values can be found in the supplementary material below.

$$\min_{z}\{\{l(z) : U(z) = x, z \in \{0,1\}^*\}. \tag{3.4}$$

where z is a program that prints string x and then halts and l is the length. The concept of Kolmogorov complexity can be used as an inference tool to find the shortest description of behavioural data. The smaller

the KC of a sequence the regular or simple it is and vice-versa. I describe the Normalized Information Distance (NID) [162], a similarity measure inspired by Kolmogorov complexity and defined as:

$$NID(x,y) = \frac{max\ \{K(x|y), K(y|x)\}}{max\ \{K(x), K(y)\}}. \tag{3.5}$$

Due to the non-computability of the NID, the NID has been re-written [163] as the normalized compression distance (NCD) by simply approximating the Kolmogorov complexity K, using a compressor Z. The NCD between two strings x and y can be defined as:

$$NCD(x,y) = \frac{Z(xy) - min\ \{Z(x), Z(y)\}}{max\ \{Z(x), Z(y)\}}. \tag{3.6}$$

Here xy is the concatenation of x and y. These strings can be documents, software, genomes or even images. The NCD takes on non-negative values in the range $0 \leq r \leq 1 + \epsilon$ with ϵ defined to take into account imperfections in the compression methods. Please refer to [163] for more details.

The NCD has been used in a variety of disciplines for different purposes, such as anomaly detection [164], gene expression dynamics [131], classification of music [165], detection and classification of computer worms and viruses as well as detecting the origin of new ones [166]. Since the NCD has been shown to work well with sequences and strings, it can be used, for example, in monitoring the behaviour of animals to know when they deviate from a previously or commonly known sequence of states, for example because of climate change [167]. It can also be used to quantify similarity in movement patterns of conspecifics across different habitats.

Species	**No of individuals**	**Habitat**	**Feeding strategy**
African buffalo (B)	5	Land	Herbivore
African elephant (E)	5(3)	Land	Herbivore
Black-backed jackal (J)	11(4)	Land	Scavenger
Burchell's zebra (Z)	2(2)	Land	Herbivore
California sea lion (sl)	15	Water	Piscivore
Galapagos albatross (A)	8	Land/Air	Piscivore
Galapagos tortoise (T)	7(4)	Land/Water	Piscivore/Herbivore
N. elephant seal (SE)	14	Water	Piscivore
Sheep & Sheep-dog (SD)	5	Land	Herbivore
Springbok (Sp)	9(1)	Land	Herbivore
White-blacked vulture (V)	4(2)	Land/Air	Scavenger

Table 3.1: **Summary of 85 individuals within 11 species used.** The code in front of the species represents the label codes used for each animals in the hierarchical clustering while the number in the bracket under the no of individuals represents the number of animals with up to one year of observational data.

3.4.2 Case study: Kolmogorov complexity as tool for classifying animal movement patterns across scales

Animals across different habitats, scales, and species are known to have different movement patterns. However, little or no study has been carried out to find out which groups of animals possess similar movement strategies across different habitats and scales. Recently in [168], the authors discussed a classification of several animals across different species into similar groups using principal component analysis on some movement metrics with hierarchical clustering. Their result suggests that all animals organizes into four distinct groups of movement syndromes namely migratory, central place, nomadic and territorial. In this case study, I analyse the movement patterns of eleven animals (Table 3.1) across different spatio-temporal scales and habitats. For this analysis, I obtained the datasets of the Galapagos tortoise (*Geochelone nigra*) [169, 170], Springbok (*Antidorcas marsupialis*) [170, 171], African buffalo (*Syncerus caffer*) [170, 172–174], African elephant (*Loxodonta africana*)(original unpublished data contributed by Miriam Tsalyuk and Wayne M. Getz)[170], Black-backed jackal (*Canis mesomelas*) [170, 175], California sea lion (*Zalophus californianus*) (original unpublished data contributed by Dan Costa)[168, 170], Galapagos albatross (*Phoebastria irrorata*) [176], Sheep (*Ovies aries*) and Sheepdog [157, 177], Northern elephant seal (*Mirounga angustirostris*) [170, 178], White-backed vulture (*Gyps africanus*) [170, 179, 180] and Burchell's Zebra (*Equus burchellii*) [181, 182]. All these datasets have the same 1 hour sampling period [168]. First, I evaluated the approach on synthetic data details of which can be found in Appendix A. Results, with both short and long movement data shows the synthetic animals can be separated distinctly into groups based on the distribution from which they sample their velocities. Then, I compared the monthly movement patterns of all the 85 animals to find similarities by computing their pairwise NCD with the gzip compressor [183] followed by hierarchical clustering of the resulting distance matrix (see Appendix A below for details). The only metric used here was distance covered every hour, which I further processed to its binary equivalent (strings of zeroes and ones). I refrained from using the turn-angle here as it is an unreliable metric considering the noisy nature of most sensors. Results (Figure 3.3) show three groups emerged: those that live on land (zebra, elephant, springbok, jackal, sheep and buffalo), those that live in water (tortoise, sea lion and elephant seal) and those that fly (Albatross and Turkey vulture). Amongst the animals that live on land, It can be noticed that there appears to be some similarity between the movement patterns of the elephant and zebra while others seem to organize into distinct groups of conspecifics. Therefore, I hypothesize that there might be a correlation between the feeding and movement patterns of animals. I also observed a small number of unexpected classifications: for example,

Vulture V_1 was classified among the animals that live in water. A possible explanation is related to the fact that the dataset is noisy. To find long-term similarities among animals movement patterns I compared approximately one year movement data of 16 animals (Table 3.1) across six different species. I selected these datasets due to their temporal length. Results (Figure 3.4) show that there might be some similarities in the long-term movement patterns of vultures and jackals which are both scavengers. This supports a hypothesis that there might be a correlation in the movement patterns of animals with similar feeding habits.

I compared this approach with the pairwise mutual information of the distance covered with respect to the two instances discussed above (see Appendix A below).

3.5 Limitations and Open Challenges

I have highlighted the potential use of information theoretic metrics in obtaining insights about animal movement and I have also showed the applications of these metrics to real animal movement data. However, I would like to underline that these methods should be applied with caution given their inherent limitations. First, the issue of missing data remains a challenging problem due to logger failure or inability to regularly obtain position fixes. In addition, it suffices to state here that care must be taken while choosing the appropriate amount of data from which inference can be made. For example, let us consider the analysis of the similarities of the complexities of animals across different taxa and spatio-temporal scales. I identified similarity only in the long-term movement patterns of jackals and vultures. This phenomenon might not have been observed if datasets of shorter length had been used. Indeed the results derived from using the measures presented in this paper should always be evaluated considering the corresponding temporal scale of the dataset.

I issue a caveat on estimating probability density functions (PDF) of continuous movement data. At the moment, most of the methods and tools available are based on the assumption of underlying normal distributions. Considering that continuous animal movement data often follows skewed (e.g., power-law or truncated power-law) distributions [184, 185], researchers employing some of the methods described here, for example the Kullback-Leibler divergence, should exercise appropriate caution while estimating the PDF of these distributions. In the present study, I took special precaution while computing the probability distribution for entropy by binning the data around several mean values of the data in the direction of skewness.

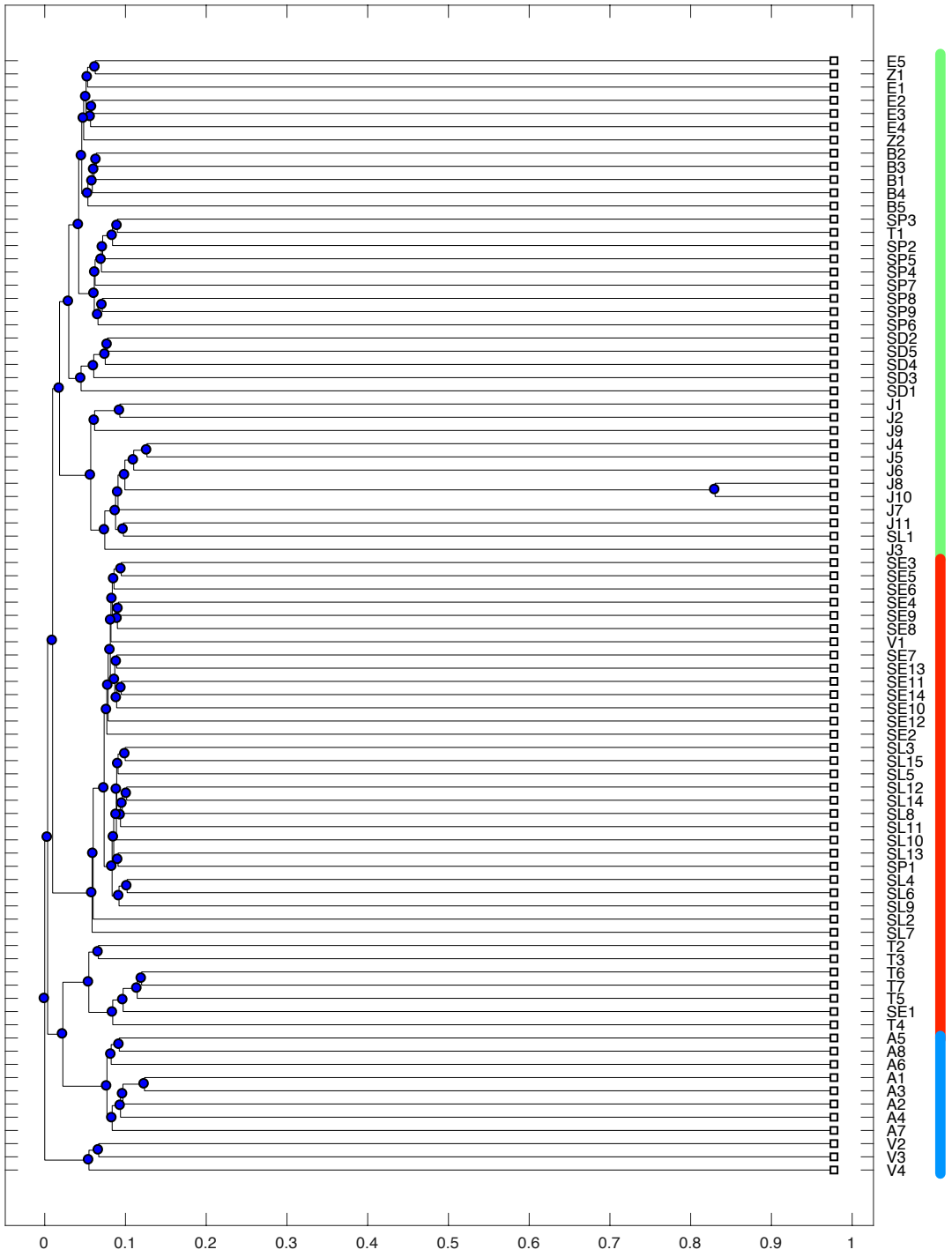


Figure 3.3: **Hierarchical clustering of the pairwise NCD of 85 animals spread across 11 species.** All the animals on average organize into three groups of those that live on land (green), those that live in water (red) as well as those who fly (blue).

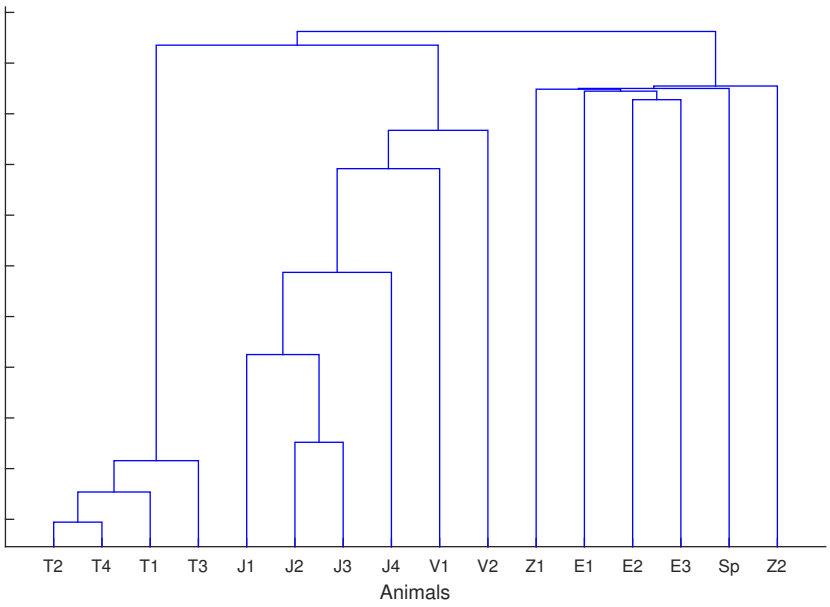


Figure 3.4: **Hierarchical clustering of the pairwise NCD of 16 animals.** These animals are spread across 6 species representing the movement patterns over a period of one year. The animals organize into three distinct groups that are correlated with their feeding patterns.

Furthermore, appropriate methods for permutation and randomization must be used while determining a threshold, for example for the calculation of the pairwise mutual information between animals in a group.

3.6 Conclusions

I have demonstrated the use of a class of non-parametric information theoretic tools for studying movement patterns of animals and have showed how they can be applied by means of several animal movement datasets. First, I have shown how Shannon entropy can be used to characterize the movement patterns of sheep with Batten disease where the distance covered every ten minutes was used as the basis for generating symbols to compute the entropy. The result shows that the Batten sheep have a lower entropy than their control counterparts. Second, I have described the use of mutual information for detecting associations in animals using pigeons as an example. Findings show that this method can be very useful in lieu of the widely used gambit of the group approach. Lastly, I have described a metric with foundations in the field of algorithmic information theory known as Kolmogorov complexity (normalized compression distance). I have used this metric to characterize the movement patterns of animals across different taxa and spatio-temporal scales

with results suggesting there might be a correlation between the feeding and movement patterns of animals. These methods provide *complementary* insights in the study of animal behaviour. In particular, they can be used to formulate new hypotheses regarding animal movement. In other words, they provide additional information about animal movement that may not be apparent using other types of analysis. This class of probabilistic methods is also usually more robust in presence of noise, which is inherent in location data.

In the next chapter, I will discuss the first application of unsupervised methods in this thesis towards quantifying abnormal movement behaviour in sheep with Batten disease.

Chapter 4

Quantifying Abnormal Movement Behaviour in Sheep with Batten Disease

In this chapter, I address the first question raised in this thesis: Can automatic methods to quantify abnormal behaviour in sheep using unsupervised machine learning techniques be developed? Here, I demonstrate in this thesis, the first use of unsupervised methods to quantify abnormal movement behaviour of sheep with Batten disease in a flock.

4.1 Introduction & Motivation

The analysis of the movement patterns of different parts of the body such as eyes [186], muscles [187] and legs [188] in humans and animals [189, 190] is important for assessing their overall welfare. Investigating abnormal locomotion patterns, however, is key towards early diagnosis of a number of neurodegenerative diseases [191]. Due to the insidious progression and terminal nature of these diseases, efficient and reliable markers are becoming increasingly important to ensure people with these neurological diseases are diagnosed early, given adequate and timely care, attention and treatment and also that the impact of such treatment can be assessed objectively. Batten disease is a rare and fatal autosomal neuro-degenerative disorder [192], char-

acterized by abnormal and involuntary movement patterns (choreoathetosis) [193], personality changes [194], loss of vision [149], and loss of muscle control (ataxia) [148] in its sufferers. It is known to be the most common form of a group of disorders called Neuronal Ceroid Lipofuscinoses (NCLs) and caused by autosomal recessive mutations in the CLN gene [195] and to be specific the CLN3 gene. The gradual progression of this disease and consequent worsening of symptoms has a profound negative impact on the quality of life of its sufferers, making them less independent with time and resulting in death in the medium term. The need to study the progression of this disease as well as develop therapeutic intervention schemes to be tested has necessitated the use of animals [196] such as mice [197] and sheep [198] for carrying out several studies in this respect. Sheep have brains that are closer to human than are those of rodents, both in terms of size and structure. They also have a longer life-span than rodents, hence their choice for this study. Of paramount interest amongst the symptoms associated with sufferers of this disease, however, is the abnormal movement [193] pattern that can be attributed to the gradual changes in the structure of the brain as the disease progresses. This abnormal movement pattern serves as a reliable marker in identifying the onset of a neurodegenerative disorder such as Batten disease. But, doing this by manual visual inspection can be laborious, time-consuming, inefficient and unreliable (Figure 4.1, right & centre). While using the distance covered in its ordinary form over the period of observation seems like a good option, it turns out that this metric is highly unreliable and uninterpretable (Figure 4.1, left & Table 4.1). There is, therefore, a need to define reliable biomarkers for automatic detection and quantification of abnormal movement patterns. Using standard anomalous identification techniques proves useful when there are two distinct populations of normal and abnormal phenotypes in a flock but becomes inadequate when there are diverse phenotypes in the population.

To this end, I propose two steps for automatically identifying sheep with unusual movement phenotypes in a flock. First, I use conventional anomaly detection methods to find abnormal movement patterns given two phenotypes in the flock and extend this approach to multiple phenotypes by using an information theoretic approach to estimate the number of groups in the flock while assuming the anomalous sheep is in the smallest group. I expatiate on these methods later below.

The rest of this chapter is organised as follows. First, I discuss the materials and methods used followed by the experiments and then the results. The chapter is concluded by highlighting the limitations of the methods used.

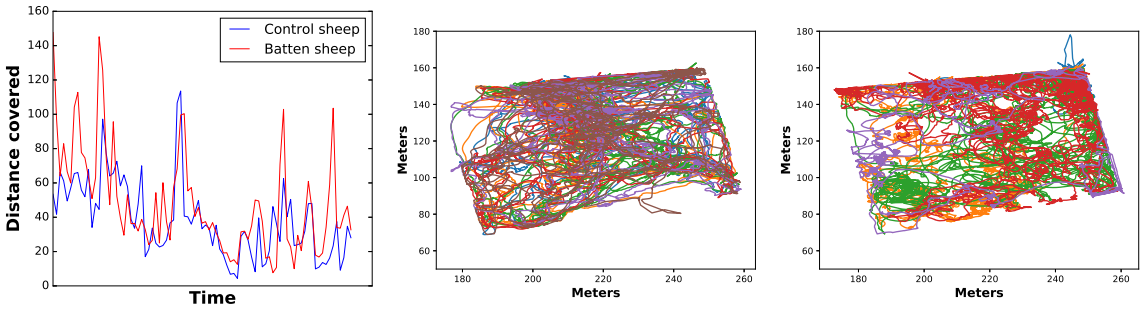


Figure 4.1: **Distance covered and trajectories of the sheep.** Left: Mean distance covered every ten minutes by the two groups of sheep on a typical day. Centre: Trajectories of the control/normal sheep. Right: Trajectories of the Batten/abnormal sheep. The mean distance covered by the two groups show this metric is highly unreliable and uninterpretable as the Batten sheep can even appear to be more active on the average than the control sheep sometimes. This also applies to visualizing the trajectories.

		Distance covered in Kilometres					
Sheep ID	Batten Disease	Day1	Day2	Day3	Day4	Day5	Day6
1	No	3.71	4.06	8.30	1.97	2.50	5.56
2	No	4.39	4.71	11.996	2.50	2.92	3.14
3	No	4.11	3.97	6.31	2.80	3.51	4.05
4	No	3.53	4.15	7.50	2.93	3.59	5.12
5	No	3.79	3.16	3.904	2.27	3.02	2.49
6	No	4.39	4.33	6.71	2.08	2.36	3.85
7	Yes	1.685	2.90	3.46	2.166	3.16	2.67
8	Yes	4.91	9.15	14.396	19.78	5.38	3.24
9	Yes	8.67	5.76	9.40	2.92	4.63	4.86
10	Yes	8.29	8.07	1.75	4.973	5.04	9.46
11	Yes	2.50	3.68	5.91	1.83	3.51	3.60

Table 4.1: **Distance covered.** Table showing the 11 sheep of interest and the corresponding distance covered in Kilometres over the course of observation during the six days of the experiment.

4.2 Materials & Methods

4.2.1 Materials

4.2.1.1 Animals

Eleven (5 Batten disease ewes carrying a mutation in CLN5 [199] and 6 healthy ewes age matched control) with mean age of 2 years were maintained at Lincoln University, New Zealand (NZ) and reared under appropriate procedures approved by the Lincoln University Animal Ethics Committee in compliance with the NZ Animal Welfare Act (1999) and in accordance with US National Institutes of Health guidelines. The sheep had water and grazing available to them throughout the period of the study.

4.2.1.2 Data Acquisition & Processing

Data for this experiment were collected for six days between 21/03/2011 and 03/04/2011. GPS data were collected using bespoke GPS loggers, details of which can be found in [200]. These were attached to the back of the sheep as shown in Figure 4.2. Previous work has shown that such harness equipment carried by sheep does not affect their locomotion [201]. Notwithstanding that sheep are slow moving creatures, GPS data at a sampling rate of 1 sample/s were collected to ensure all forms of interesting movement patterns by the sheep were captured. The data were further processed using the open-source RTKLib and Geographic Library [202] to obtain Cartesian co-ordinates with respect to a local projection. All missing data were interpolated between the last and next seen co-ordinates using a Kalman smoother, the parameters of which were optimised using expectation maximisation [150][1]. The datasets were all roughly 24 hours in length. Each dataset comprises a phase in which loggers were attached to each sheep in a holding pen; a phase in which the sheep were herded into the field; a phase in which the sheep were left to roam across the field; and a final phase in which the sheep were herded back to the holding pen to have the logging device removed and re-charged. All computations were made based on the data obtained when the sheep were left to wander in the field on their own as this represents the period in which their natural behaviour was most likely to be observed (see Table 4.3 for the number of hours extracted). The eleven sheep were kept together during the first three days of the experiment, and were mixed with another group of sixty nine sheep that had also been kept together as well for the remaining three days. The sixty nine sheep consisted of younger animals in a mixture of unaffected sheep and other sheep at different stages of progression with respect to a variety of

[1]The dataset had already been preprocessed this way. See the cited literature.

medical conditions (see Table 4.2). For the experiments, I extracted the portion of the data in which the animals were on the field until at least one of the GPS loggers attached to the sheep failed due to battery exhaustion. All software was created using Matlab (R2016a).

	G1	G2	G3	G4	G5	G6
First 3 days	-	-	-	-	5	6
Last 3 days	5	8	11	17	23	16

Table 4.2: **Summary of the representative genotypes in the flock across the six days of the experiments.** Where G5 & G6 represents the control and Batten sheep respectively.

Day	Duration of observation (hours)
1	18
2	18
3	22
4	13
5	16
6	19

Table 4.3: **Duration in hours of the length of data extracted for the purpose of this work.**

4.2.2 Methods

4.2.2.1 Discretizing the Trajectory

For each of the sheep in the flock, I obtained Cartesian trajectories from the GPS traces and partitioned them into periods of ten minutes each. The total distance covered on a per second basis was computed and summed up over each period of ten minutes per sheep across all observation windows. Ten minutes was chosen as it was considered long enough time to observe any significant changes in behaviour and also because it minimises the effect of GPS noise.

4.2.2.2 Detecting and Aggregating Outliers

There are several methods to detect outliers[2] in a dataset such as [203] in which a density-based metric was used, [204] in which the K-nearest neighbour concept was used and [205] in which a distance metric was used. I used three fully unsupervised methods discussed below because most of the methods listed above are either not fully unsupervised with requirements for training and test set or they are less parsimonious with requirements for setting a number of parameters. The K-nearest neighbour approach, for example, requires

[2]The unusual distance(s) covered in the population every ten minutes has/have been assumed as outliers where unusual here means the distance(s) covered different from the majority using some similarity metrics.

Figure 4.2: **Loggers on sheep.** Sheep with GPS loggers attached to their back. Research has shown that such harness equipment does not affect their locomotion [201].

the number of nearest neighbours to be specified; which influences the result of the analysis. The local outlier factor approach, in addition to requiring the specification of the number of nearest neighbours to be used, also requires a training and test set. The three methods used are therefore Gaussian mixture model (GMM), K-means and Gaussian based models because they require little or no parameters to select compared to the those described above. I expand on these methods further below. The number of outliers were summed over the entire window of observation. Here, it is the case that the control sheep were always in the majority and the number of clusters is two.

4.2.3 Models

4.2.3.1 Gaussian Mixture Model (GMM)

This is an unsupervised probabilistic model that assumes all data points are generated from a mixture of Gaussian distributions. This approach incorporates uncertainty and assigns each data point to a cluster with some measure of probability. I assumed that the cluster with the smallest membership is the anomalous group.

4.2.3.2 K-means

The K-means algorithm is another unsupervised clustering algorithm that finds K groups in data by iteratively assigning each data point to one of K groups based on its distance to a centroid until convergence. This is a hard assignment compared to GMM with a soft assignment. As with GMM, the cluster with the smallest membership is assumed to be the anomalous group.

4.2.3.3 Gaussian Model-Based

This approach assumes data is generated from a Gaussian distribution and that all data that are more than three standard deviations away from the distribution mean are outliers. I examined two instances by using the distance covered every ten minutes as well as the log of this data. This is because the distance covered does not assume a perfect Gaussian distribution.

$$\textbf{outliers} < \mu \pm 3\sigma \; or \; \mu \pm 3\sigma < \textbf{outliers} \tag{4.1}$$

Where μ is mean and σ the standard deviation.

4.3 Experiments & Results

4.3.1 Synthetic Experiments

Two types of synthetic experiments reflective of the two scenarios considered in this chapter[3] were carried out.

4.3.1.1 Experiments I

Three main groups of experiments were conducted to evaluate the performance of the methods for quantifying abnormal movement behaviour. To evaluate the performance of the four methods, I applied them to a set of synthetic data representative of possible sheep velocities. First, ten bins of bivariate Gaussians each with 10000 random points were generated where:

[3]See section 4.3.2 below for these two scenarios.

$$mean\ (\mu) = \begin{bmatrix} 0 & 0 \\ 0.0001 & 0.0001 \\ 0.001 & 0.001 \\ 0.01 & 0.01 \\ 0.1 & 0.1 \\ 1 & 1 \\ 1.5 & 1.5 \\ 2 & 2 \\ 2.5 & 2.5 \\ 3 & 3 \end{bmatrix}$$

and covariance matrix (CM):

$$CM = \begin{bmatrix} \mu/4 & \mu/4 \\ \mu/4 & \mu/4 \end{bmatrix}$$

I created 11 agents who randomly sampled their velocities uniformly from the bins at any instant of time[4] (see sample trajectories in Figure 4.3). The normal agents were designed to be in the majority[5], and also sampled their velocity from the same bin while the abnormal agents sampled from different bins. This was designed to create a situation reflective of behaviour of sheep in the flock of interest here where the control sheep have a somewhat uniform behaviour while the Batten sheep have abnormal behaviours of different types. There are however three instances of this scenario here, when the abnormal agents sampled from just one bin, when they sampled from two different bins and when they sampled from three different bins. The number of the abnormal agents was designed to be less than or equal to five[6]. To account for temporal variability in the duration of abnormal movements, the duration during which the abnormal agents sampled from the abnormal distribution was varied from $0 : t/4 : t$ where $t = 600$ seconds[7]. The performance of the four methods were evaluated first, based on the duration of abnormal movements using two metrics, the classification accuracy as well as the false positives. And then evaluated, based on the number of agents

[4]Summed over 144 instances representing every ten minutes for 24 hours.

[5]Implying that at no point in time were they less than six.

[6]A random number generated between 1 & 5 to select the number of abnormal agents to sample from a different bin with respect to the normal ones each time window.

[7]Used because ten minutes was also considered for the actual sheep movement data.

involved in the abnormal movements across several time windows with respect to the two metrics over a thousand iterations.

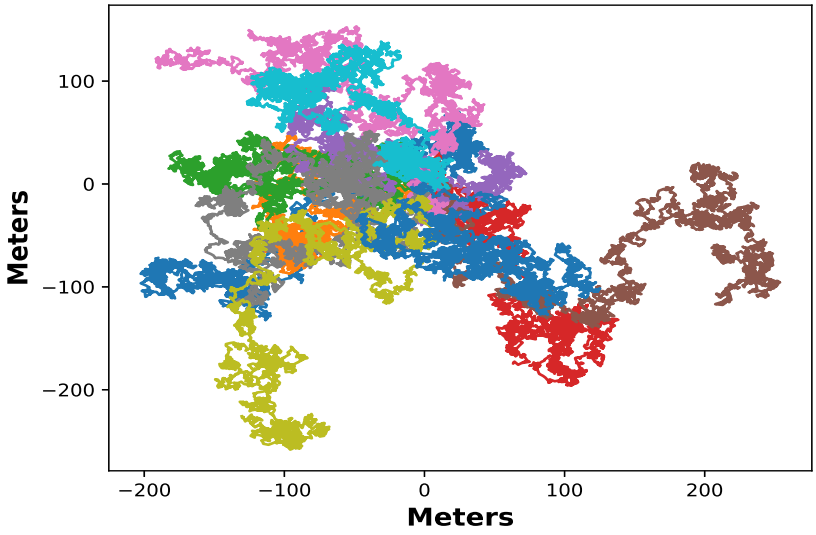


Figure 4.3: **Samples of the synthetic trajectories.**

Results in Figure 4.4 show GMM and K-means achieve better accuracies compared to the Gaussian model based[8] across all three contexts. However, while GMM have small false $+ves$ that are fairly constant or decreases across the three contexts, other models displayed some inconsistencies with respect to this metric. For instance, the false $+ves$ for K-means is the highest in the first context and drops significantly after as the diversity of unusual behaviour increases. Whereas, for the Gaussian model based methods, while the false $+ves$ appear lower initially, this starts to increase as the diversity of abnormal samples increases. Overall, the results suggest that the clustering based methods may be more robust given diversity in the abnormal behaviour but with accuracy decreasing and vice versa for the Gaussian based models most especially when the unusual behaviour is short in duration.

4.3.1.2 Experiments II

Experiments II are similar to Experiments I described above. The only difference being that I used only two methods[9] together with the proposed algorithm in section 4.3.2.2 below. I investigated the optimum number of clusters ranging from 1 to 5 and evaluated the performance of the algorithms as before in terms of

[8]Accuracy decreases with time and number of individuals involved.
[9]The clustering methods.

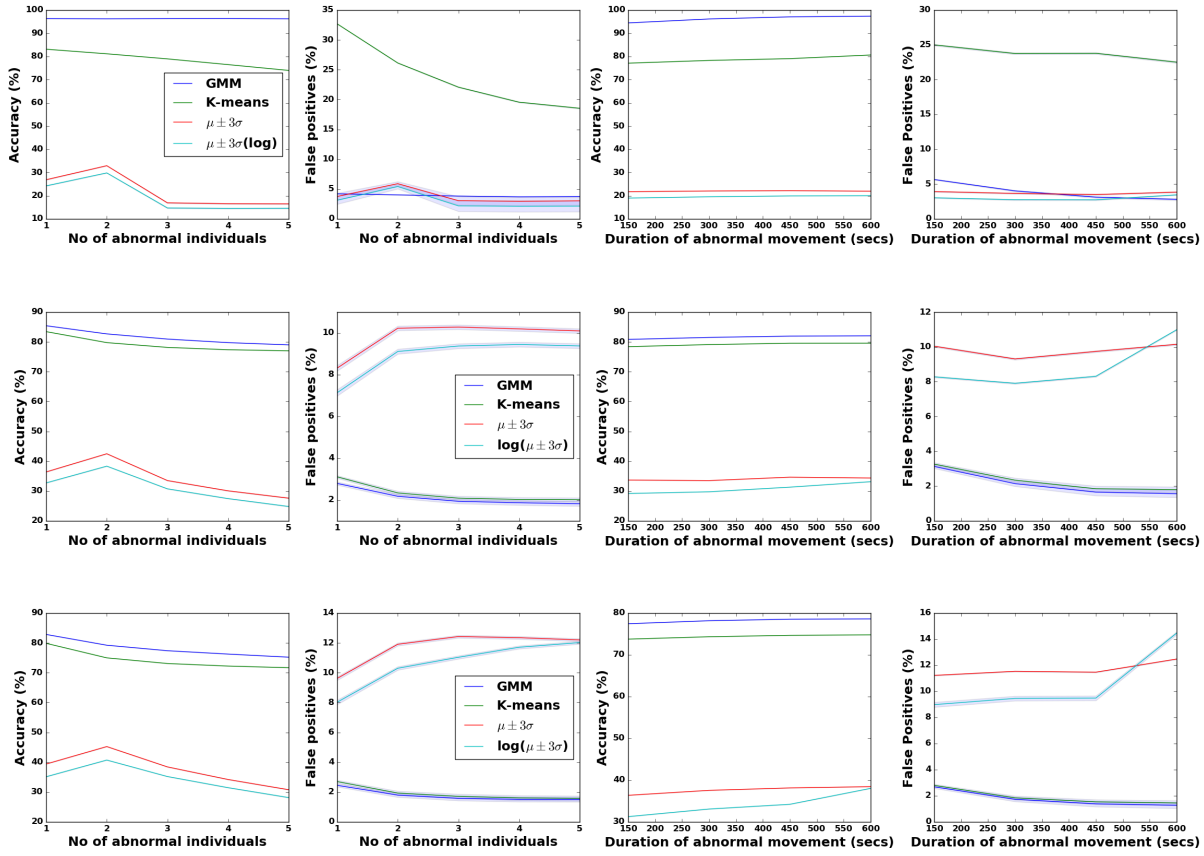


Figure 4.4: **Synthetic Experiments I**. Top row: When unusual movement is sampled from one distribution. Middle row: unusual movement is sampled from two different distributions. Bottom row: unusual movement sampled from three different distributions. While the accuracies appear consistent across the three contexts between models, decreasing slightly for the clustering based models and increasing for the rest. The false $+ves$ generally decreases as the diversity of the abnormal samples increases. Individually, the accuracies and false $+ves$ with respect to the variation in the number of abnormal individuals appears to have approximately the same behaviour decreasing slightly (which is probably due to the fact the clustering algorithms may be struggling to estimate the appropriate parameters of the clusters as the number of abnormal individuals grow towards that of the normal ones) for the clustering based algorithm while peaking at two (likely optimal threshold number, enough to separate abnormal and normal behaviour given the diversity in the behaviour of abnormal individuals) and then decreasing afterwards as the number of abnormal individuals increases. On the other hand, with respect to individual model performance in relation to the duration of abnormal movement, the clustering based algorithms accuracies increases and its false $+ves$ decreases both slightly as the duration of abnormal movement increases while for other models, the accuracies and the false $+ves$ increases as the duration of abnormal movement increases. Finally, because logarithm functions grow slightly faster than linear ones, the accuracies and false $+ves$ of the log version of the gaussian model based method can be seen to rise faster than its version without the log.

the classification accuracy and false positives.

Results, in Figure 4.5 show the identification accuracy increases as the duration of abnormal movement as well as the number of agents involved increases on the average. This applies to the false $+ves$ as well. Across the three contexts, the false $+ves$ appear higher for K-means compared to GMM and vice-versa for the accuracies. This suggests GMM might be more suitable for this task due to the soft nature of its class assignment.

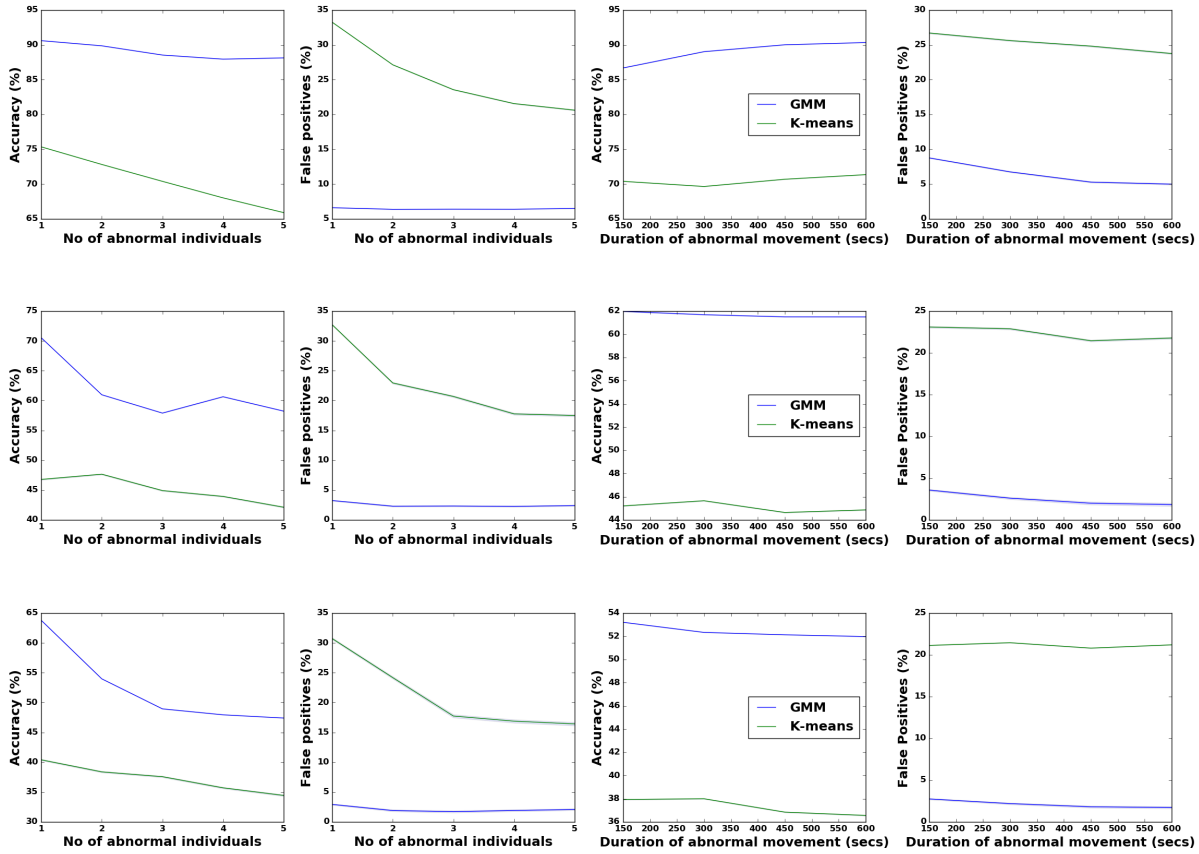


Figure 4.5: **Synthetic experiments II**. Top row: When unusual movement is from one distribution. Middle row: When unusual movement is sampled from two distributions. Bottom row: When unusual movement is sampled from three different distributions. The accuracies and false $+ves$ can be seen to decrease as the diversity of unusual movement behaviour increases.

4.3.2 Experiments with Sheep Mobility Dataset

4.3.2.1 Quantifying Abnormal Movement Behaviour

I used the model and procedures described above, considering only the eleven sheep of interest. It can be observed that all the sheep with the natural mutation for Batten disease have at least one outlier every day over the period of observation across the four methods whereas this is not the case for the control sheep group. In addition, it can also be seen that the bulk of the outliers accumulated by the flock over the period of observation across six days belong to the Batten sheep group, as seen in Figure 4.7, even though they represent a minority (5) of the flock. Also, among the sheep with Batten disease, it can be observed that there is a considerable variation in the total number of outliers accumulated all through the period of observation. This suggests that these sheep have different trajectories in terms of progression of this disease and possible variation in reaction as well. Furthermore, it can be seen that there are no specific patterns in the magnitude of the outliers seen. They can be as low as 10 metres and can be as high as 500 metres using the gaussian based model, for example (see Figure 4.6), suggesting that these abnormal movements can be random and also a function of how different sheep react to the disease.

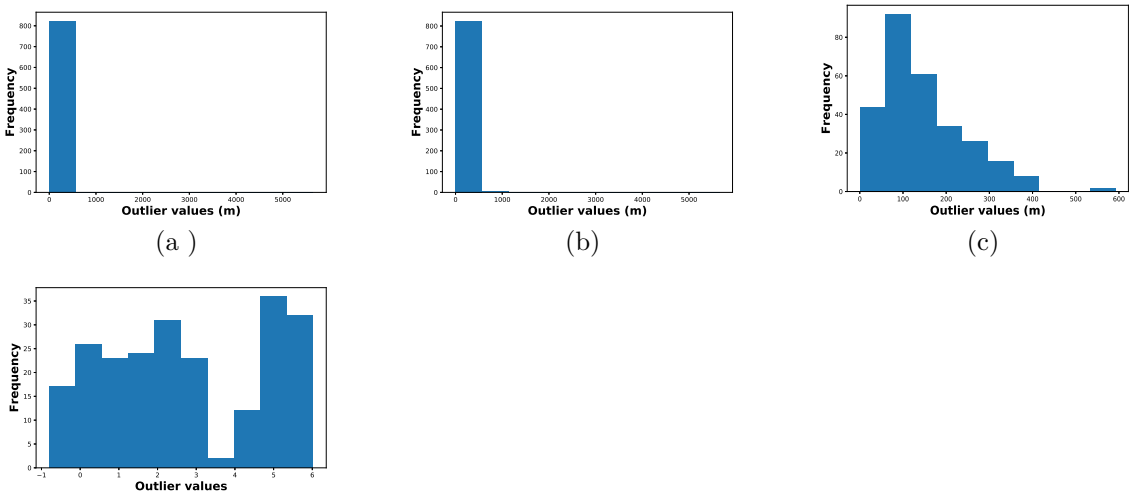


Figure 4.6: **Distribution of the outliers** (a) GMM (b) K-means (c) Gaussian based model (d) Gaussian based model using log of data.

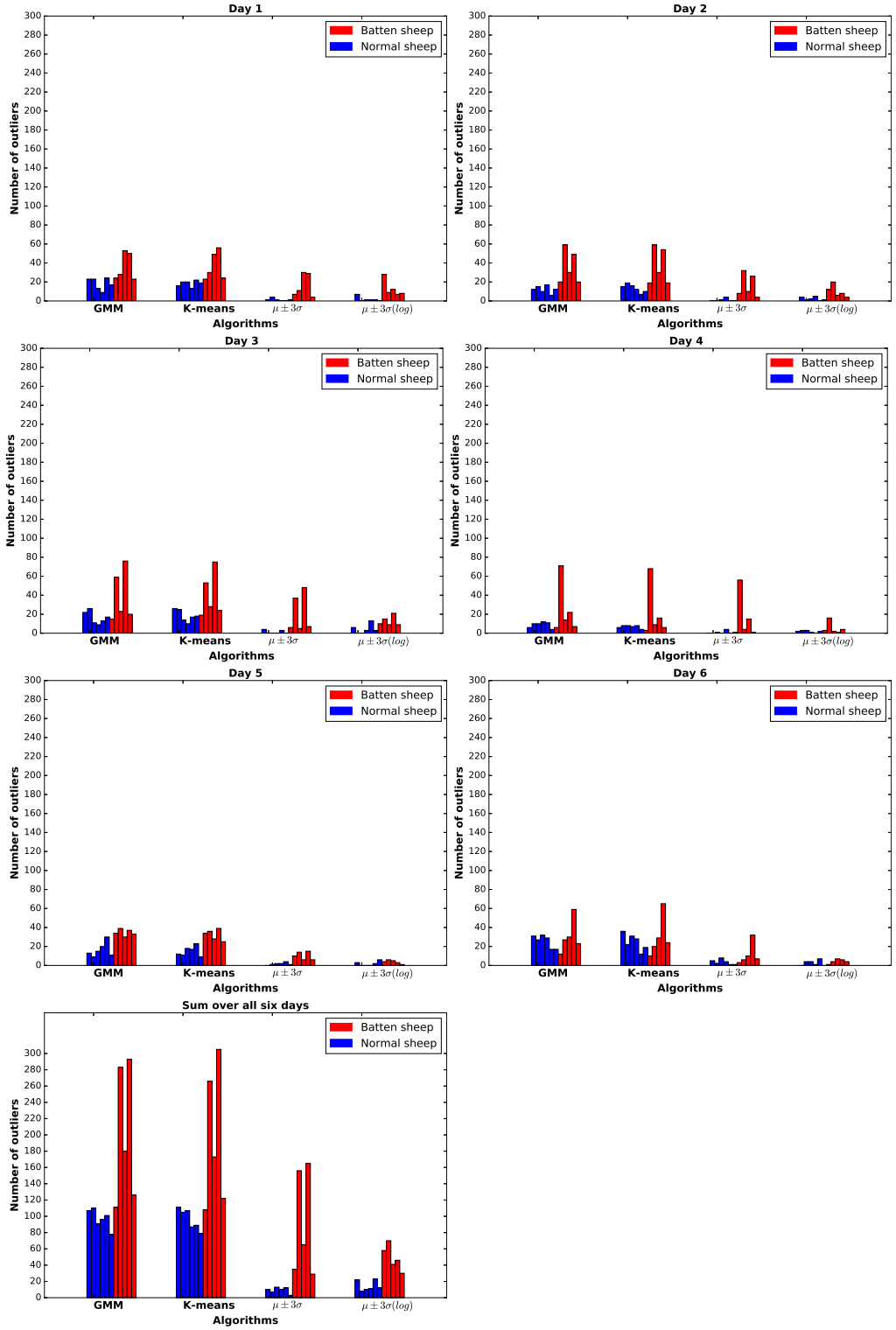


Figure 4.7: **Quantifying abnormal behaviour in a small flock.** Bar chart showing the number of outliers detected over six days when the sheep of interest were considered as a flock without others in their environment. It can be seen with all four methods that the sum total of the number of outliers in the Batten sheep is greater thus confirming its importance as a biomarker. The IDs of the sheep is in the order listed in Table 4.1 above where each bar represents each sheep of interest.

4.3.2.2 Quantifying Abnormal Movement Behaviour in a More Challenging Context

I investigated the possibility of quantifying abnormal behaviour in a more challenging context by considering not only the movement of the Batten sheep relative to their control group but to the whole of the flock during the last three days of the experiment. The previous methods were used initially but results showed they are ineffective because, in a larger group, it is expected that there will be a number of sub-phenotypes in the flock, rather than just two. Therefore, I proposed an alternative method. The Akaike information criterion (AIC) was used to select the optimum number of sub-groups[10] [206] based on the distance covered every ten minutes in the dataset and then this was used to cluster the sheep into groups using both the Gaussian mixture model as well as the K-means method (Algorithm 1). I hypothesized that the Batten sheep will most likely be in the sub-group with the smallest membership. These methods were evaluated on the dataset of movement activities in the last three days of the experiment. Results, show that at least three of the sheep with Batten disease have a conspicuous number of outliers relative to other sheep of interest in this experiment (see Figure 4.8). This can even be higher given significant observation period for the experiments.

$$accuracy = \frac{Predicted\ significant\ \&\ insignificant\ pairwise\ relationships}{Ground\ truth\ significant\ \&\ insignificant\ pairwise\ relationships} \tag{4.2}$$

Algorithm 1

Input: Movement trajectories of flock (X), **Output:** abb
$X_m \leftarrow$ Segment movement trajectory
for each segment m in X_m:
    $N_m \leftarrow$ AIC to determine optimum no of clusters in X_m.
    $IDs \leftarrow$ Cluster X_m using N_m & log identities of sheep in the smallest cluster.
    Append IDs in *abb*
end
return abb

Learning how significant the difference in the abnormal behaviour of the sheep in this study is important for example, in assessing the efficacy of therapeutic interventions. To further evaluate the performance of all methods, I proposed an equation (4.2) that computes the ratio of how significant the unusual phenotype quantified here is between the control and the transgenic sheep and how insignificant it is in between the control sheep with respect to the ground truth. Here, it is expected that the pairwise test should be insignificant between the control sheep and significant between the control & transgenic sheep[11]. I used the Wilcoxon signed-rank test with $p \leq 0.05$ as threshold to determine significance. Results for the first approach

[10] Between 1 and 20.
[11] See example of this result in Table 4.4 and the rest in Appendix B.

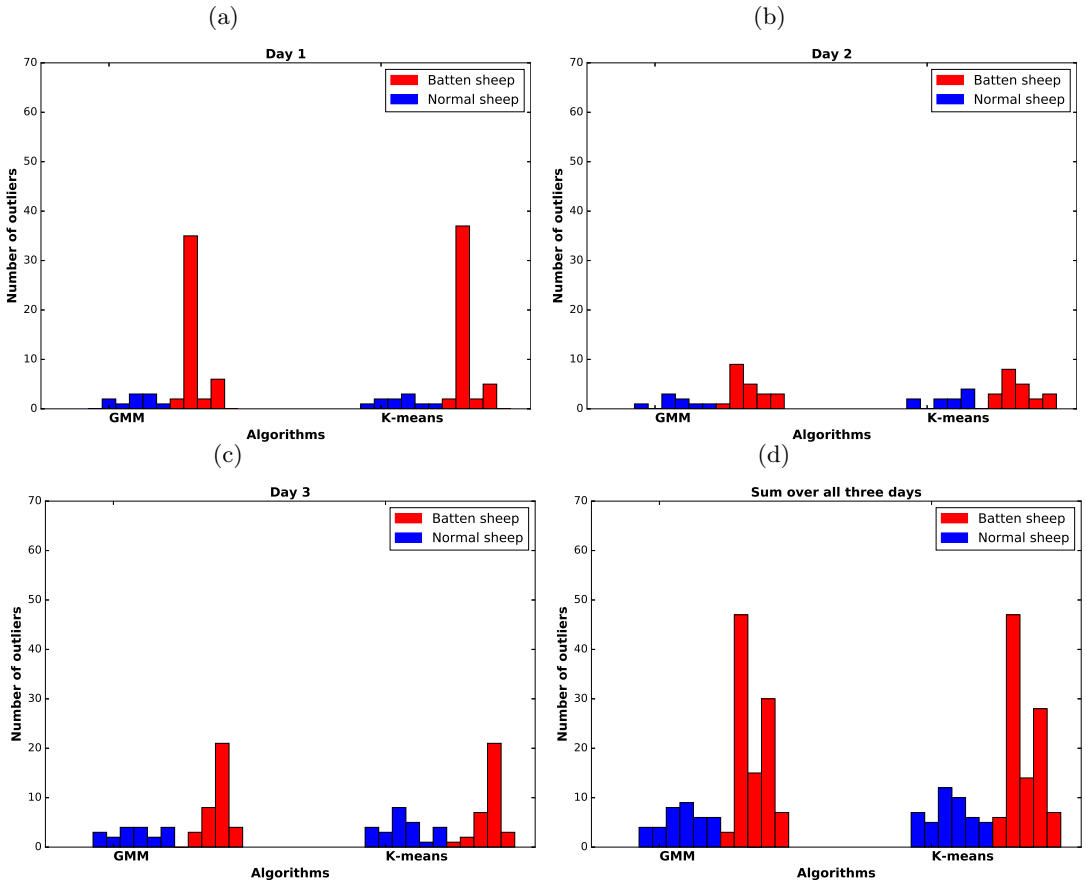


Figure 4.8: **Quantifying abnormal behaviour in a bigger flock.** Bar chart of outliers for the two groups of sheep when the movement patterns of other sheep were taken into considerations on the (a) first day (b) second day (c) third day and (d) the sum over three days. It can be seen that at least three of the Batten sheep have a conspicuously higher number of outliers relative to the rest. The IDs of the sheep is in the order listed in Table 4.1 where each bar represents each sheep of interest.

		Sheep ID										
Sheep ID	Batten Disease	1	2	3	4	5	6	7	8	9	10	11
1	No	-	NS	NS	NS	NS	NS	NS	NS	NS	S	NS
2	No	NS	-	NS	NS	NS	NS	NS	NS	NS	S	NS
3	No	NS	NS	-	NS	NS	NS	NS	NS	NS	S	NS
4	No	NS	NS	NS	-	NS	NS	NS	NS	S	S	NS
5	No	NS	NS	NS	NS	-	NS	NS	S	NS	S	NS
6	No	NS	NS	NS	NS	NS	-	NS	S	S	S	S

Table 4.4: **Pairwise test of significance results using GMM.** Where S = Significant when p-value $\leq$ 0.05, and NS = Not significant when p-value $>$ 0.05. See Appendix B for more results. Note: Sheep with IDs 7 to 11 are sheep with Batten disease.

Methods	GMM	K-means	$\mu \pm 3\sigma$	$\mu \pm 3\sigma$ (log)
Accuracy	68.33%	65%	83.33%	61.67%

Table 4.5: **Significance influenced classification.** Accuracy across the two groups of sheep when they were considered irrespective of others in their environment. See Appendix B for confusion matrices.

(see Table 4.5) where the sheep of interest were considered without regard for others in their environment show the difference in behaviour is not statistically significant between the control sheep across the four methods confirming they have a somewhat uniform behaviour (see Table 4.4 and Appendix B for more results). On the other hand, difference in behaviour is significant between the two groups of sheep and between the Batten sheep, with the gaussian model based approach performing best. One plausible explanation for this given the synthetic experiments is that, there are several of the Batten sheep with diversity in their abnormal behaviour at any instant of time. With respect to evaluation considering other sheep in the flock, I observed no significant difference within the control sheep and some significance between the two groups (see Table 4.6 and Appendix B) with the GMM method performing better than K-means, which is consistent with the simulations. In addition, the confusion matrices in Figures 4.9 and 4.10 show that, while all the methods appear to be adept at recognising the insignificant relationships they struggle in recognising the significant relationships with the Gaussian model based appearing to perform best in this regard in the first scenario and the GMM in the second scenario.

Methods	GMM	K-means
Accuracy	66.67%	60%

Table 4.6: **Significance influenced classification.** Accuracy across the two groups of sheep when they were considered with respect to other sheep in their environment.

	Significant	**Non-Significant**
Significant	0.37	0.63
Non-Significant	0	1

(a) GMM

	Significant	**Non-Significant**
Significant	0.43	0.57
Non-Significant	0.13	0.87

(b) K-means

	Significant	**Non-Significant**
Significant	0.67	0.33
Non-Significant	0	1

(c) $\mu \pm 3\sigma$

	Significant	**Non-Significant**
Significant	0.23	0.77
Non-Significant	0	1

(d) $\mu \pm 3\sigma$(log)

Figure 4.9: Confusion matrices for the four methods in the first scenario considering the pairwise significant and non-significant relationships.

	Significant	Non-Significant
Significant	0.33	0.67
Non-Significant	0	1

(a) GMM

	Significant	Non-Significant
Significant	0.2	0.8
Non-Significant	0	1

(b) K-means

Figure 4.10: Confusion matrices for the second scenario considering the pairwise significant and non-significant relationships.

4.4 Limitations & Conclusion

In this work, I have proposed several unsupervised methods towards quantifying abnormal movement behaviour and, by extension, identifying sheep with these characteristics in a flock being one of the numerous symptoms associated with sheep with Batten disease. I achieved this by computing the distance covered by every member of the flock within a period and looking for outliers in the population. This work is not without its limitations: the methods proposed here would not scale in situations where the abnormal sheep are in the majority as the methods are designed for instances where the unusual behaviour is in the minority in the population. A class of methods focused on learning collective movement behaviour [207] will be useful here where unusual statistics[12] of the collective behaviour with respect to some known ground truth of normal behaviour can provide useful insights with respect to the welfare of the flock or herd. Whilst the sample size (11 sheep) is small during the first three days of the experiment, I devised a method that works well in a challenging environment as experienced in the last three days of the experiment. The implication of this work transcends identifying sheep with abnormal movement patterns as the methods can be used to monitor the health of individuals or agents that co-exist in groups and have the potential to develop some form of movement disorder during their lifetime. This is a step towards quantifying the progression of a number of neurodegenerative diseases using sheep as subjects. In the future, these methods will be applied to another dataset of transgenic sheep that are expected to display subtle symptoms of Huntington's disease with the aim of identifying the onset and progression of the condition.

Overall in this chapter, I have demonstrated the use of unsupervised methods towards quantifying abnormal movement patterns in sheep with Batten disease. In the next chapter, I will discuss the first application of a supervised approach in this thesis using a recurrent neural network towards forecasting avian migration patterns.

[12]Unusual statistics here means behavioural measurements that differs a lot from what is generally known.

Chapter 5

Forecasting Avian Migration Patterns

In the previous chapter, I discussed the application of unsupervised methods in quantifying abnormal movement behaviour. In this chapter, I address the second question raised in this thesis: Can a bidirectional recurrent neural network model designed to forecast migration states of Turkey vulture birds at a finer granularity[1] across different continents perform better than alternatives? And also by extension, what environmental features are most indicative of different migration states with respect to this avian species? Here, I show the first application of supervised methods in this thesis by using a deep bidirectional recurrent neural network augmented with an auxiliary task for instance in forecasting migration states in light of sustainable efforts geared towards conservation and reducing biodiversity loss.

5.1 Introduction & Motivation

Reducing biodiversity loss is fundamental towards achieving the 2030 agenda of the United Nations sustainable development goals [208]. With 6% of avian life functionally extinct and 21% currently threatened [209], biodiversity loss with respect to global avifauna deserves serious attention [210]. The importance of birds in nature's ecosystem cannot be overemphasized. They have multiple important roles to play in the ecosystem as they serve as pollinators by processing and transporting plant seed from one geographical area to another [211]; help control the population of insects [212], which in turn helps improve agricultural yield; play host to many host-specific parasites such as lice [213] as well as supply eggs as food to humans (domesticated species).

[1] Most common being migration intensities but movement every one hour will be considered here.

However, an estimated 100 million to 1 billion birds are known to die annually during migration via collisions with human made critical infrastructures such as wind turbines, buildings and high tension lines [214, 215]. These collisions can sometimes be due to partial blindness caused by excessive light [216] at night on several of these infrastructures or the inability to see them when flying at night leading to high impact collisions. Optimal management of such infrastructures is therefore key towards reducing avian biodiversity loss across the world, and this would be enabled by having better prediction of animal movement and migration patterns.

While previous works [66] and [67] have focused on forecasting the intensity of spring migration and autumn migration respectively, the onset and the end of these migration states are of utmost importance towards ensuring a concerted approach to reducing biodiversity loss for example when animals migrate between continents. In addition, I argue that forecasting the intensity of migration cannot be guaranteed to be optimal for managing critical infrastructures as different species of birds potentially have different migration dates along different routes. This suggests that having an independent model for each migrating avian species, most importantly the threatened species would be valuable to handle the specific nuances associated with their migration patterns. To this end, I consider a much more difficult task of forecasting different avian migration/movement states (breeding, wintering, fall and spring migration) at a finer granularity and to be specific on an hourly basis.

In this chapter, the task of forecasting migration states is considered. More specifically, I use publicly available datasets of Turkey vulture (*Cathartes aura*) movement collected over multiple years at a resolution of one hour. I train a deep learning network consisting of Bidirectional-Recurrent Neural Networks[2] by casting the prediction problem as a supervised learning task with a cross entropy loss where I aim to forecast the onset of four movement states corresponding to breeding, fall migration, non-breeding and spring migration. The network is augmented by adding an auxiliary task to help stabilize training and improve generalization by forecasting the longitude[3]. Experimental results of forecasting at different temporal intervals in advance show good accuracies can be achieved even when forecasting up to one week in advance. I compare this approach to a variety of baselines and empirically show that it outperform them most of the time on a variety of experiments.

The rest of this chapter is organized as follows. Next, I give a brief overview of the avian species used in this work followed by the model and the datasets as well as the associated preprocessing. This is followed by

[2]Used to add more context for prediction. For instance an ordinary recurrent neural network scans data in one direction (left to right) whereas a bidirectional version in addition also scans from right to left with the concatenation used for prediction.

[3]The most informative migration feature.

the experiments and results sections in that order. I conclude this chapter by highlighting the limitations of the methods used.

5.2 Background: The Turkey vulture

The Turkey vulture (*Cathartes aura*) represents the most widely distributed scavenger in the world with global population in excess of five million spread across the two American continents and the West Indies [217]. Gregarious in nature, it has been regarded as a partial migrant [218] as some populations winter in North America while others move as far as South America en-masse (with population in excess of 2000 previously recorded [219]) during the winter months while returning to breed during spring. The latter population is considered in this work. Breeding, including incubation responsibilities, is usually carried out by both parents spanning late spring to early winter months. They are primarily scavengers with a good sense of smell that helps them locate carcasses. By eating carcasses, they help make the environment cleaner and prevent the spread of diseases. The main predators of Turkey vultures are Owls, hawks and eagles, with their eggs serving as food for raccoons and opossums. While the population of these birds is stable and they are not endangered, the approach proposed in this chapter can be extended to the endangered avian species given the relevant migration data.

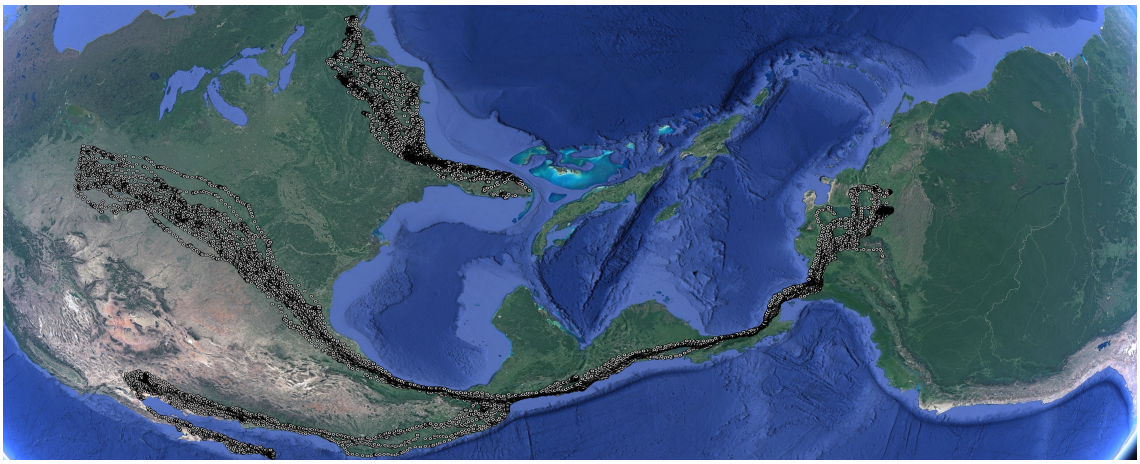


Figure 5.1: **The North America continent on the left and South America on the right as seen from Google Earth.** A significant population of the Turkey vulture in North America are known to migrate (trajectories in black) to the south on the flight for survival to escape the unfavourable weather conditions as winter approaches.

5.3 Datasets

Turkey vulture movement datasets [220] and [221] as part of an ongoing study were collected with the aid of GPS satellite transmitters. This dataset was collected to assess the relationship between the wing size and shape as well as the characteristics of the migratory habits of the turkey vulture. Information collected in the dataset include the event id, latitude, longitude, timestamp, binary information on the visibility of the event on the Movebank search map, ground speed, heading, height, migration state, individual and tag local identifiers among others. The birds used are representative of five populations covering different breeding areas in central Canada, western and eastern USA, south-west Argentina and the Falkland Islands. Overall, about 127 birds were tracked with the focus on a subset of these birds that are migratory[4]. A significant portion of the data for the migratory birds were collected at the temporal resolution of one hour, with others at three hours. I discarded the points collected at three hour intervals so as to be able to work at a more fine grained resolution with more data-points.

5.3.1 Migration States

There are four movement states in the datasets, see [221] for more information. In this work, I seek the onset of these states which are described in detail below. Here, onset is regarded as the first recorded instance of a migration or movement state.

End of spring migration/Onset of breeding: Period after spring migration when breeding starts. During breeding, birds mate, give birth and raise fledglings in preparation for fall migration. This is usually between March and April every year.

End of breeding/Onset of fall migration: Start of migration to the wintering ground usually in South America. This is usually between September and October every year.

End of fall migration/Onset of wintering: Arrival to the wintering ground in South America which heralds the start of the wintering season. This is usually around October each year.

End of wintering/Onset of spring migration: Start of migration back to North America when the weather must have improved enough for breeding to start. This is usually between March and April every year.

[4] 27 in number.

5.3.2 Environmental & Weather Data

The movement tracks were annotated with environmental data with the aid of the Env-DATA Track Annotation Service [222]. Twenty-six features were collected: population density, atmospheric water, downward ultraviolet radiation at the surface, plant canopy surface water at surface, incident solar radiation, elevation, dew point temperature, snow evaporation, water vapour concentration, snow albedo, temperature parameter in canopy conductance at surface, surface solar radiation downwards, albedo, soil temperature, surface solar radiation, maximum temperature, evaporation, ice temperature, land surface temperature night, land surface temperature day, snow temperature, surface thermal radiation, ten metre wind gust, sunshine duration, surface thermal radiation downwards and soil water content.

5.3.3 Data Preprocessing

I obtained the trajectories of three birds out of all the birds (16 in number) in the datasets corresponding to the ones with migration state labels and fewest missing data (Rosalie, Mac & Morongo) with breeding sites in Washington (USA), Saskatchewan (Canada) and Oregon (USA) respectively. Also the birds above have respectively four years of movement trajectories collected with size[5] 28613, 11900 GPS points consisting of one full year of movement activities and another half a year of breeding and fall migration activities as well as 20753 GPS points consisting of movement activities around a period of 3 years. I divided these trajectories into 75% for training and the rest for testing (Tables 5.1, 5.2 and 5.3). All missing data in the environmental & weather data were replaced with the last observed points in the sequence.

Data	Breeding	Fall Migration	Wintering	Spring Migration
Training	0.5046	0.0495	0.3795	0.0664
Test	0.5123	0.0625	0.3639	0.0614

Table 5.1: **Class distribution (Rosalie).** With respect to breeding, fall migration, wintering and spring migration for both training and test dataset.

Data	Breeding	Fall Migration	Wintering	Spring Migration
Training	0.47	0.26	0.15	0.12
Test	0.28	0.6	0.02	0.1

Table 5.2: **Class distribution (Morongo).** With respect to breeding, fall migration, wintering and spring migration for both training and test dataset.

[5]Number of GPS points.

Data	Breeding	Fall Migration	Wintering	Spring Migration
Training	0.422	0.2848	0.1379	0.1553
Test	0.4017	0.5983	-	-

Table 5.3: **Class distribution (Mac).** With respect to breeding, fall migration, wintering and spring migration for both training and test dataset.

5.4 Problem Formulation & Model

5.4.1 Problem

Given a time series $\{X_{\leq t}, y_{mt}\}_{t=n}^{T}$ where $X_{\leq t} \in \mathbb{R}^d$ and y_{mt} is a multi-dimensional vector representing input features ($X_{\leq t} = X_{t-d}, ..., X_t$)[6] and discrete migration states respectively at each time-step t where $m \geq 2$, the goal is to estimate $\{y_{m(t+k)}\}_{t=n}^{T}$ given $\{X_{\leq t}\}_{t=n}^{T-k}$ where k is the temporal period in advance (forecast horizon) with respect to the migration state of interest.

5.4.2 Model

5.4.2.1 Stacked Bidirectional GRU (Bi-GRU)

The gated recurrent unit (GRU) [223] network is a variant of the recurrent neural network (RNN) that learns to integrate temporal information over sequential data using a combination of gates. Its advantage over the conventional RNN is its ability to overcome the vanishing gradient problem. However, the GRU only integrates information from one end of the data to another. To add better context to a RNN, the Bidirectional RNN [224] was proposed. The Bi-GRU thus, is a version of the Bidirectional RNN using the GRU and works by training two GRUs where one is trained on the forward hidden input sequence $\overrightarrow{h}$ and the other on the reversed copy of this sequence $\overleftarrow{h}$ using two hidden layers to compute the output sequence y_t given input $\mathbf{x}$. Given that the GRU equations in [223] can be implemented by the function G, the Bi-GRU can then be implemented via:

$$\begin{aligned} \overrightarrow{h_t} &= G(\overrightarrow{h}_{t-1}, \mathbf{x}_t) \\ \overleftarrow{h_t} &= G(\overleftarrow{h}_{t+1}, \mathbf{x}_t) \\ y_t &= W_{\overrightarrow{h}} \overrightarrow{h_t} + W_{\overleftarrow{h}} \overleftarrow{h_t} + b \end{aligned} \tag{5.1}$$

[6] d is the duration of the temporal context relevant for the prediction task.

This approach has been demonstrated to yield better results. For example it was used previously to fill in missing values [225] and hence the choice in this regard given geolocation data almost always have this characteristic. With stacked Bi-GRU, the expressiveness of the Bi-GRU is improved. Given N layers of the Bi-GRU, the hidden and output sequences are updated as follows:

$$\begin{aligned} \overrightarrow{h}_t^n &= G(\overrightarrow{h}_{t-1}^n, \overrightarrow{h}_t^{n-1}) \\ \overleftarrow{h}_t^n &= G(\overleftarrow{h}_{t+1}^n, \overleftarrow{h}_t^{n-1}) \\ y_t &= W_{\overrightarrow{h}} \overrightarrow{h}_t^n + W_{\overleftarrow{h}} \overleftarrow{h}_t^n + b \end{aligned} \tag{5.2}$$

Where $h_t^{n-1} = \mathbf{x}$ when $n = 1$.

5.4.2.2 Auxiliary Tasks

To help improve training, ensure stable learning dynamics, as well as learn a robust representation that generalises to test data, auxiliary tasks have been used to provide even denser training signals to support the main task [226]. Given all the input features are continuous, the network loss is therefore augmented with the mean square error (MSE) loss (single task) where I aim to forecast the longitude coordinate which is the most informative feature relevant to the main task (see Section 5.6).

5.4.2.3 Training Objectives

At training time, I seek the parameters θ_1 and θ_2 that minimizes the cross entropy loss, $\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_{mi}, f(X_i;\theta_1))$ for forecasting the migration state and the MSE loss, $\frac{1}{N}\sum_{i=1}^{N}(f(X_i;\theta_2) - y_{li})^2$ for forecasting the longitude coordinates respectively. The overall training objective therefore is given by:

$$\underset{\theta_1,\theta_2}{argmin}\ E(\theta_1,\theta_2) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_{mi}, f(X_i;\theta_1)) + \frac{1}{N}\sum_{i=1}^{N}(f(X_i;\theta_2) - y_{li})^2 \tag{5.3}$$

Where X_i, y_{mi} and y_{li} represents the input features, migration states and longitude coordinates respectively and f is a function approximator such as a neural network.

5.5 Experiments & Procedures

I used 3 layers of Bi-GRU as described in Section 4.3 with a dropout layer for the main task, bidirectional LSTM networks work equally well in this case and either can be used. Given the size of the dataset, I trained

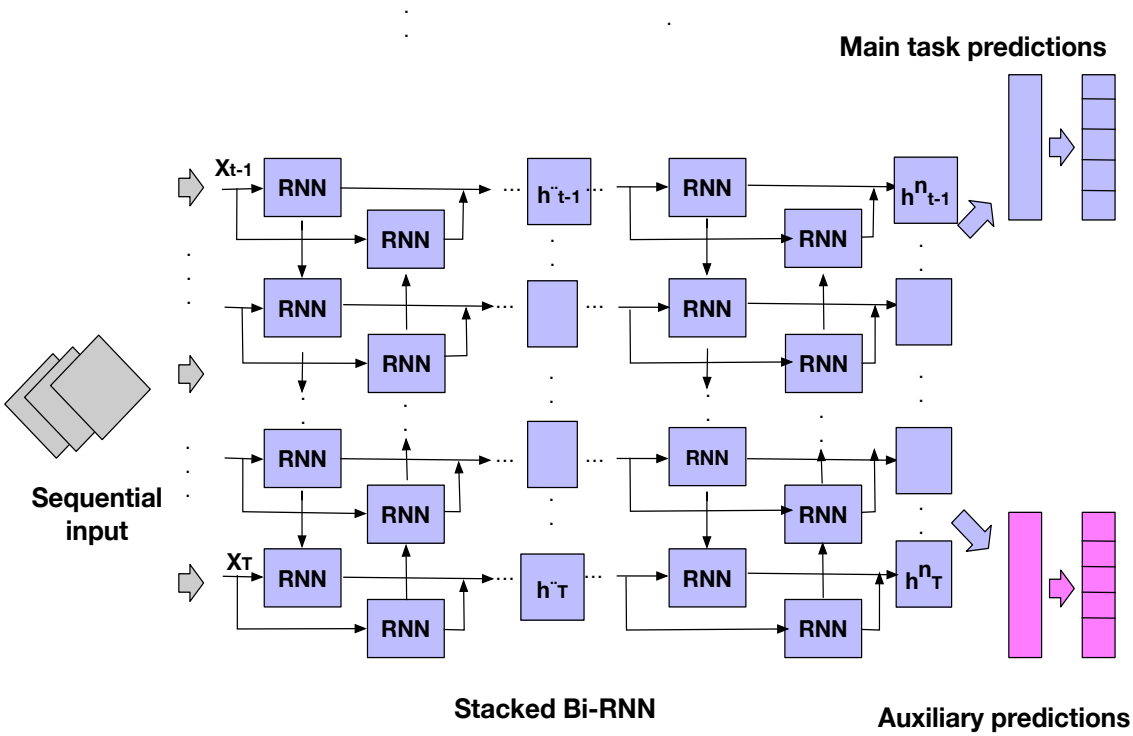


Figure 5.2: **Architecture of the proposed framework.** The main predictions represent the actual task of interest which is the forecasting of the migration states while the auxiliary predictions represent the forecasting of the longitude.

the network over 20 epochs three times with early stopping after two epochs without improvement in the loss. The models were evaluated on the test sets for the same amount of time and the average reported. I evaluated all models using accuracy and F1 scores. No significance tests were carried out.

5.5.1 Thresholding

Predictions of machine learning models can only be reliable when the accuracy is perfect. Given the nature of real world data, it is unlikely any model would work perfectly thus rendering these models tricky for real world critical applications. Calibration methods such as Platt scaling [227] and isotonic regression [228] are useful in reducing this problem. However, in this case they are ineffective for two reasons. First, [229] argued that deep learning models with softmax output trained with cross-entropy loss are prone to errors. Second, the distribution of the training data is hardly the same as test data due to several factors affecting the migration states. While the obvious approach after calibrating is to select a threshold of the class probability scores given the ROC curve, the prediction around the boundary surrounding the transition from one migration

state to another in this case can be problematic. To get around this problem, I instead propose an algorithm (Algorithm 2) that is model dependent and leverages a sequence of thresholds to predict the migration states. The intuition behind the algorithm is that I decide a migration or movement state has started or ended if there is a continuous sequence of a threshold value of the probability scores. The temporal difference between the end of the first instance of this sequence and the ground truth averaged over the number of years for each migration state in the training data is then used in the test data probability distribution to predict the migration states. I used a probability threshold score $th_{max} = 0.71$[7] and sequence length l of 24 below. See Figures 5.8, 5.9 and 5.10 for extensive comparisons of different sequence lengths using this threshold.

Algorithm 2

Input: Training dataset $(X_{\leq t}, y_{mt})$
Output: thr
$Pr(y_{mt}|X_{\leq t}) \leftarrow$ *equation* 4.3
$th_l \leftarrow$Sequence of threshold th_{max} of length l (first instance)
for each unique m in y_{mt}:
  for each year in (y_{mt}):
    $thr\{m\} \leftarrow thr\{m\} + (t_{onset\ of\ m(ground\ tr.)} - t_{th_l})$
  end
  $thr\{m\} \leftarrow$ thr{m} / No of years in $(X_{\leq t}, y_{mt})$
end
return thr

5.5.2 Baselines

I compared the proposed approach to several baselines including deep and non deep learning methods[8]. These baselines to the best of my knowledge are competitive for time-series classification tasks.

- **Logistic regression (LR):** I used a multinomial variant with a LBFGS solver where the usual sigmoid logistic function is replaced by a softmax function and the class probability which sums up to one is computed as a function of the inputs.
- **Adaboost (AB):** The Adaboost algorithm uses an ensemble of weak learners combined and weighted to arrive at a model's prediction whilst iteratively improving on instances misclassified by previous classifiers [77]. Used with a hundred estimators.
- **Support Vector Machine (SVM):** This algorithm categorizes data according to optimal hyperplanes that maximizes the distance separating the classes of interest given the training data. Used with a one

[7]This is the maximum probability seen in the training set using the model.
[8]Grid search used to estimate some of the parameters.

versus one strategy.

- **Deep Neural Network (DNN):** I used a feed-forward network with two layers and 256 neurons each, dropout = 0.2, softmax layer at the output, relu activation and Adam opimizer [230].
- **Recurrent Neural Network (GRU)** A GRU network with 1 layer, 50 cells, dropout = 0.2, softmax layer, tanh activation and Adam optimizer [230]. I trained the neural networks given the procedures described for the proposed approach above. LSTM was tried as well and gives just the same results as the GRU.

5.5.3 Experiments

I designed several experiments described in detail below to answer several questions related to migration with respect to this avian species given the dataset I possess.

- **Q1:** What environmental and movement factors are most indicative of migration patterns? To answer this, I quantified the mutual information between the continuous environmental and movement features and discrete migration states using the approach proposed in [159]. I computed the mutual information for all the three birds mentioned above and used the average to estimate this.
- **Q2:** How does the proposed approach compare to a variety of baselines over several forecast horizons? Also, to what extent do the predictions deviate from the ground truth with respect to the migration states described in section 5.3.1? To answer Q2, I designed series of experiments where I aimed to forecast the migration states one day, three and seven days in advance given the proposed approach. For the latter part of this question, I used only the bird with the biggest trajectory size and the least amount of missing data to minimize uncertainty as much as possible and also ensure all the migration states of interest are well represented.
- **Q3:** How is the forecasting accuracy affected when there is a significant difference in the migration dates encountered in the training set compared to the test set? To answer this, I queried the date of the onset of all the migration states in the training and test datasets and compare these with the results from Q2 above. As in the latter part of the previous question, I used the bird with the biggest trajectory size and the least amount of missing data to minimize uncertainty as much as possible and also to ensure all migration states are adequately represented.
- **Q4:** Finally I aim to ask, is migration restricted to a specific period of the day? To answer this, I

queried the time of all the migration states described above in the test and training datasets to examine if a pattern can be observed.

5.6 Results & Discussion

I discuss the results and answer to several questions asked above in the order listed here.

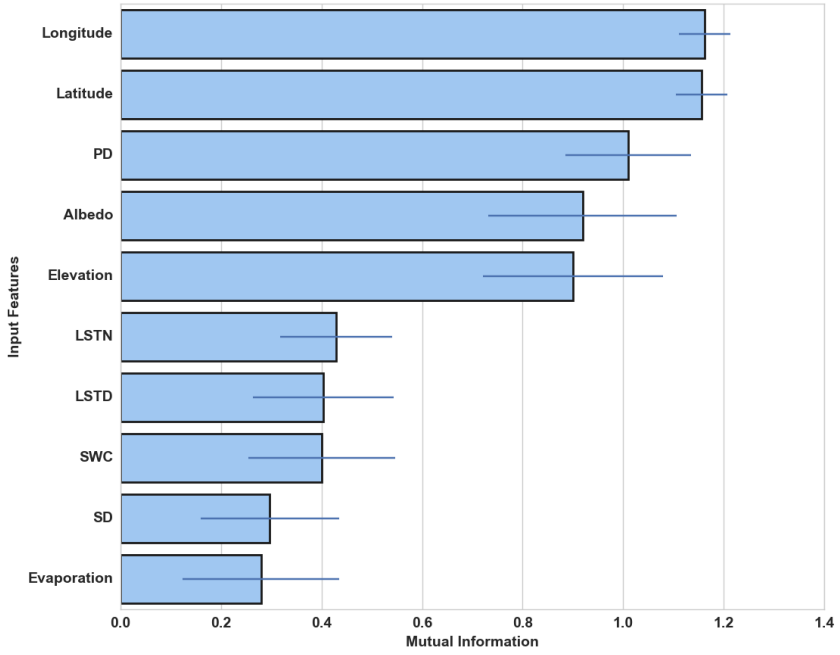


Figure 5.3: **Mutual Information between the top ten most informative input features and the migration states.** It can be seen that outside of movement coordinate features, human population density, albedo and elevation are great indicators of migration states. PD: Population Density, LSTN: Land Surface Temperature Night, LSTD: Land Surface Temperature Day, SWC: Soil Water Content, SD: Sunshine Duration.

- **A1:** The results of the experiment described in Q1 are shown in Figures 5.3 & 5.4. The top ten most relevant features show that, apart from the movement coordinates, population density, albedo, elevation, temperature and soil water content represent the greatest indicators of different migration states in this bird. This is a result consistent with a previous study [66]. I also observe from the time-series plots that the onset of fall and spring migration is preceded by a drop in temperature and soil water content on both continents. Furthermore, I notice the bird winter and breed in areas where human population density is low. This has significant implications for conservation, as wintering and breeding grounds

are lost to human development, the start of the corresponding migration states can be affected as the birds seek new grounds for these activities. In addition, the elevation of flight can be seen to be low during breeding as the birds focus on production of offspring with abundant resources, and high during wintering as they move from one place to another in search of carrions and warmer climes. Also worthy of mentioning is the albedo, a measure of light reflected from the surface of the earth. It decreases and then increases prior to autumn migration and vice versa prior to spring migration. This is a pattern in alignment with change in activities by humans as well as plants and animals. There is high albedo, for example, when temperature drops due to inactivity and low albedo as the temperature improves.

- **A2:** Results in Tables 5.4, 5.5 and 5.6 as well as Figures 5.5, 5.6 and 5.7 show that, when the dataset is big and of high quality for example as obtained in the Rosalie movement dataset, the proposed approach can be seen to outperform a variety of baselines demonstrating its effectiveness in adding more context via the backward and forward integration of information as well as using the added auxiliary task[9]. The proposed approach also outperform the baselines on the prediction of the underrepresented states when longer horizons are considered. In addition, results in Table 5.7 also show I can forecast relatively very well at different intervals of one day, three days and seven days in advance. Furthermore, ablation studies show that adding several layers of the Bi-GRU in addition to the auxiliary task helps improve result. When the datasets are either not big enough, or not of high quality (more missing data) or all the migration states not well represented, the proposed approach have almost similar performance compared to its ablation[10], performing better than the ablation sometimes and the ablation performing better some other times[11].

- **A3:** Results in Tables 5.4, 5.7 and 5.8 show the proposed approach perform not too bad when the difference was large, see for example when the onset of breeding was delayed significantly in the test data.

- **A4:** Results in Tables 5.8, 5.9 and 5.10 show that migration can start or end within a window that lies between evening time and early hours of the day UTC time. Practical interventions can therefore be implemented around this window to ensure minimal infrastructure downtime.

[9]This comes at a cost though as the recurrent models are computationally intensive with the amount of time required to train these models growing as they get deeper. On the other hand, the classical machine learning models are very easy to train with little compute resources. More specifically, it takes about three days for each forecast horizon considered for the most complex recurrent network used compared to just about five minutes for the classical machine learning methods over the ten iterations considered.

[10]Model without auxiliary task.

[11]But overall, the performance of the proposed approach and its ablation without the auxiliary task appear to be better than other baselines.

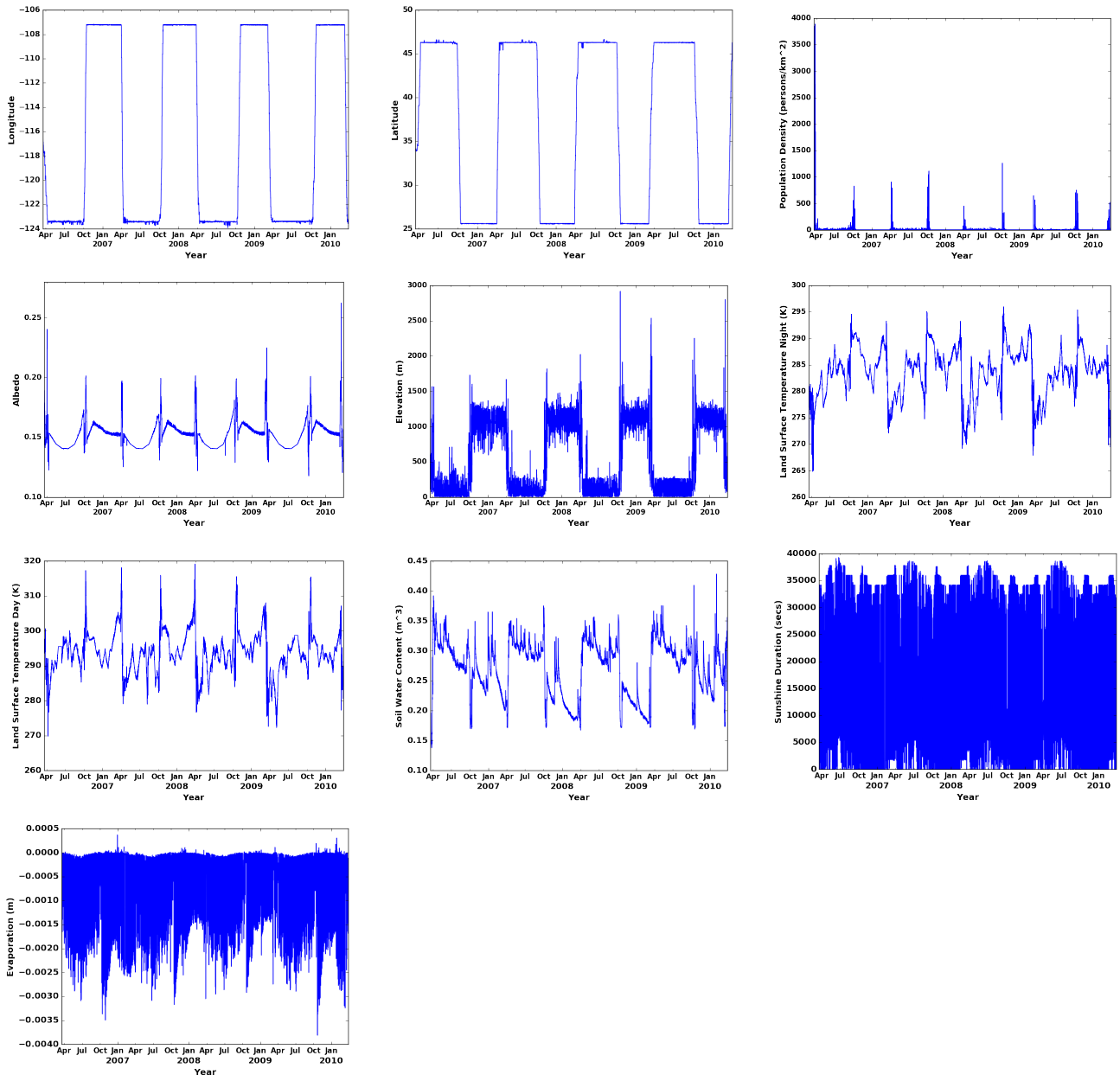


Figure 5.4: **Time series plot.** Of the top ten (for brevity) most informative features across both the training and test datasets.

No of days in advance	Performance Metric	Models							
		LR	AB	SVM	DNN	GRU (1-L)	Bi-GRU (1-L)	Bi-GRU (3-L)	Bi-GRU (3-L, A. task)
1	Acc.(%)	74.83	93.01	51.14	51.11	83.33±3.3	80±1.4	93±0.47	**93.67±0.47**
	F1 (Breeding)	0.82	**0.99**	0.68	0.68	0.92±0.03	0.88±0.01	0.96	0.96
	F1 (Fall M.)	0	0	0	0	0	0	**0.85±0.05**	**0.85±0.05**
	F1 (Wintering)	0.76	**0.99**	0	0	0.84±0.04	0.81	0.96	0.97
	F1 (Spring M.)	0.25	**0.65**	0	0	0	0.07±0.09	0.5±0.03	0.48±0.1
3	Acc.(%)	74.32	55.39	51.18	51.11	81.67±2.49	80.66±6.34	**94.33±0.47**	**93.33±0.9**
	F1 (Breeding)	0.82	**0.97**	0.68	0.68	0.9±0.02	0.88±0.06	**0.97**	0.96
	F1 (Fall M.)	0	0	0	0	0	0.08±0.12	**0.83±0.06**	0.76
	F1 (Wintering)	0.75	0	0	0	0.84±0.05	0.81±0.1	**0.98**	**0.98**
	F1 (Spring M.)	0.24	0.16	0	0	0	0.11±0.12	**0.67±0.08**	0.47
7	Acc.(%)	68.46	59.73	51.38	10.64±6.72	82±0.82	87.2±0.86	**93±2.16**	**94±1.4**
	F1 (Breeding)	0.76	0.87	0.68	0	0.92±0.01	0.93±0.02	0.97	**0.98**
	F1 (Fall M.)	0	0.36	0	0.05±0.05	0	0.31±0.12	0.85	**0.86**
	F1 (Wintering)	0.68	0.44	0	0.11±0.15	0.82	0.92±0.03	0.96±0.01	**0.97±0.02**
	F1 (Spring M.)	0	0.19	0	0.09±0.06	0.04±0.06	0.2±0.15	0.37±0.26	**0.41±0.29**

Table 5.4: **Performance comparison (Rosalie).** Proposed approach with and without the auxiliary tasks can be seen to outperform all baselines across the three experiments most of the time. L represents layer(s) & A auxiliary. Note: The recurrent networks appear to have the same performance across the three time horizons but would decrease given sufficient time in advance of prediction. The inconsistencies associated with other models such as AB is due to the way it learns. All models were evaluated over longer horizons with more runs to further evaluate the performance of the models (see Figures 5.5-5.7). The figures in bold correspond to the best performing model.

No of days in advance	Performance Metric	Models							
		LR	AB	SVM	DNN	GRU (1-L)	Bi-GRU (1-L)	Bi-GRU (3-L)	Bi-GRU (3-L, A. task)
1	Acc.(%)	37	66	28	40.33±18.26	76.33±2.62	61.67± 8	**80.33±0.47**	**81.33±2.49**
	F1 (Breeding)	0.46	**0.9**	0.44	0.23±0.16	0.79±0.06	0.63±0.07	0.86	0.89±0.01
	F1 (Fall M.)	0.33	0.66	0	0.45±0.32	0.84±0.02	0.68±0.09	**0.87**	**0.87±0.03**
	F1 (Wintering)	0	**0.4**	0	0	0.11±0.15	0.05±0.06	0.2±0.05	0.2±0.07
	F1 (Spring M.)	0	0.26	0	0.05±0.07	0.41±0.04	0.05±0.08	0.49±0.06	**0.55±0.06**
3	Acc.(%)	38	80	28	40±28.28	79.33±1.7	80.33±2.87	**82±1.63**	**81.67±2.87**
	F1 (Breeding)	0.46	**0.87**	0.43	0	0.86±0.02	0.87±0.04	**0.87±0.01**	0.83±0.03
	F1 (Fall M.)	0.35	0.84	0	0.5±0.35	0.86	0.87±0.02	0.87±0.01	**0.89±0.02**
	F1 (Wintering)	0	0	0	0	0.11±0.12	0.09±0.06	0.08±0.11	0.04±0.05
	F1 (Spring M.)	0	0.16	0	0.05±0.07	0.03±0.05	0.14±0.12	0.59±0.09	**0.62±0.08**
7	Acc.(%)	35	56	27	13±18.38	79±5.1	78.67±4.99	**84.67±1.24**	**84±2.16**
	F1 (Breeding)	0.44	0.38	0.43	0	0.86±0.08	0.84±0.07	**0.86±0.02**	**0.86**
	F1 (Fall M.)	0.3	0.69	0	0.20±0.29	0.850.04	0.85±0.03	**0.9**	**0.89±0.02**
	F1 (Wintering)	0	0	0	0	0.2±0.15	0.18±0.13	**0.26±0.19**	**0.29±0.13**
	F1 (Spring M.)	0	0.11	0	0.05±0.07	0.05±0.07	0.09±0.12	**0.71±0.08**	0.65±0.06

Table 5.5: **Performance comparison (Morongo).** Results show the deep bidirectional recurrent network with and without the auxiliary task perform better than all other baselines.

No of days in advance	Performance Metric	Models							
		LR	AB	SVM	DNN	GRU (1-L)	Bi-GRU (1-L)	Bi-GRU (3-L)	Bi-GRU (3-L, A. task)
1	Acc.(%)	52	59	40	22±4.32	73.33±8.2	65±3.5	**89.33±0.47**	**88.33±6**
	F1 (Breeding)	0.6	0.68	0.57	0.29±0.2	0.8±0.02	0.71±0.02	**0.91±0.02**	**0.89±0.04**
	F1 (Fall M.)	0.45	0.65	0	0.17±0.23	0.7±0.1	0.62±0.05	**0.91**	**0.9±0.06**
3	Acc.(%)	52	60	40	25.67±20.43	73.67±7.9	72.67±4.78	**91.67±6.6**	**86±2.16**
	F1 (Breeding)	0.58	0	0.57	0	0.77±0.1	0.76±0.07	**0.93±0.04**	0.89
	F1 (Fall M.)	0.44	0.75	0	0.42±0.3	0.75±0.1	0.76±0.04	**0.92±0.07**	0.88±0.02
7	Acc.(%)	48	80	39	35.33±4.5	75.76±1.7	69.67±3.86	**86.67**±3.4	**85.66±2.62**
	F1 (Breeding)	0.57	0.78	0.56	0.23±0.23	0.83±0.03	0.75±0.04	**0.87±0.01**	**0.87**
	F1 (Fall M.)	0.37	0.81	0	0.35±0.24	0.8±0.02	0.68±0.04	**0.87±0.04**	**0.87±0.03**

Table 5.6: **Performance comparison (Mac).** Results show the deep bidirectional recurrent network with and without the auxiliary task perform better than all other baselines. There seems to be no difference between the proposed approach and its variant without auxiliary tasks due to reasons I attribute to the size of the test and training data.

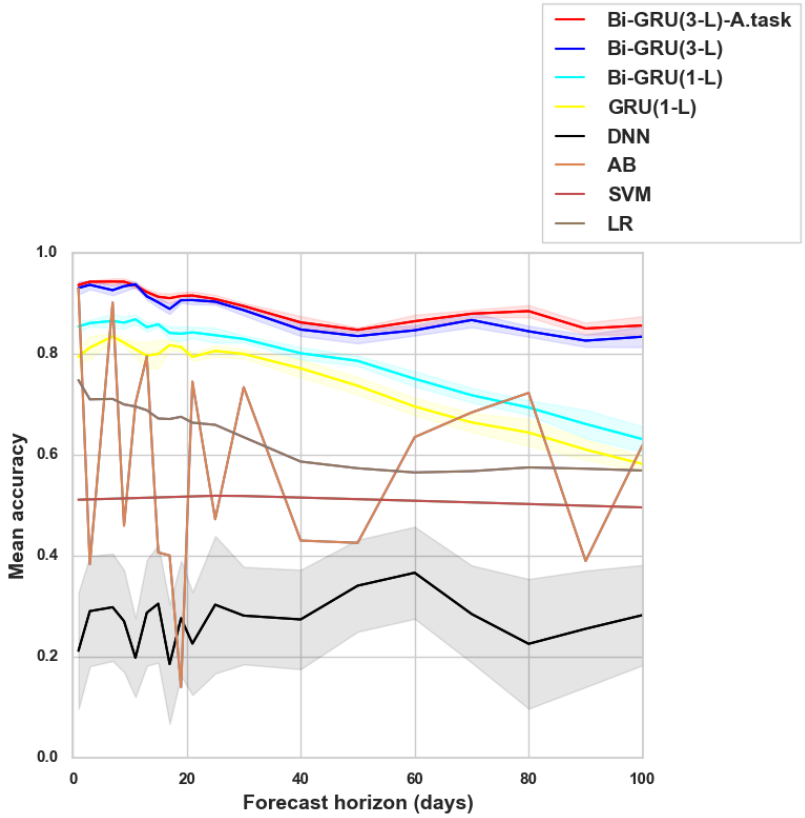


Figure 5.5: **Performance comparison (accuracy) over longer horizons (Rosalie).** Proposed approach can be seen to outperform all baselines. It can also be seen that the classical machine learning models except Logistic regression were unable to model the temporal relationship inherent in this data as the performance of these models does not degrade with time as compared to the recurrent models thus providing highly unreliable results. Note: Here, the mean and confidence intervals were computed over ten runs of the experiments as supposed to three in Table 5.4.

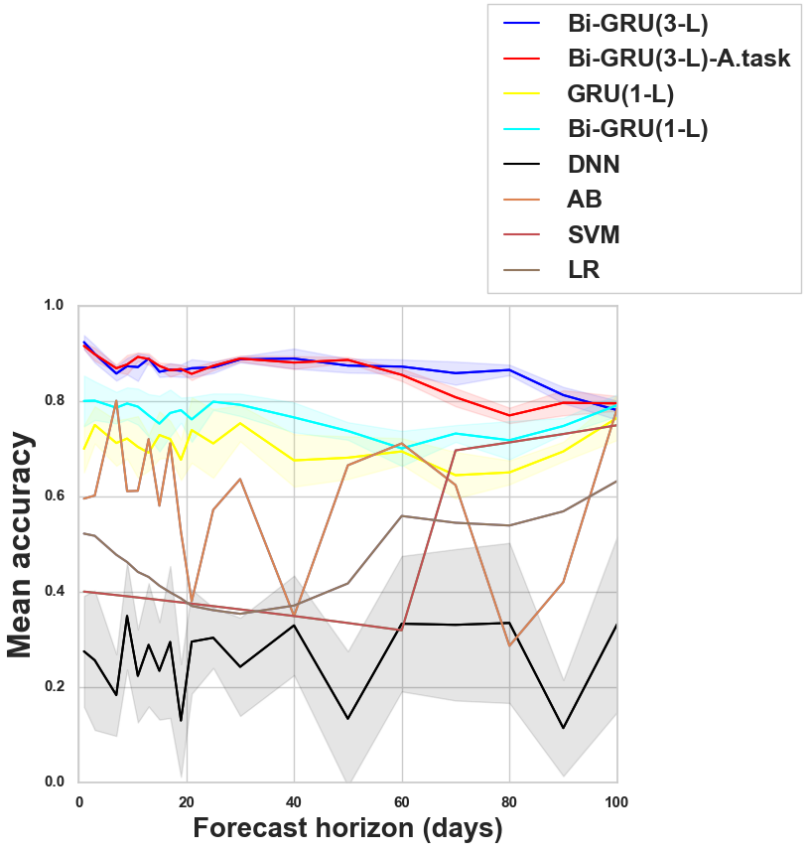


Figure 5.6: **Performance comparison (Mac) over longer horizons.** It can also be seen that the classical machine learning models are unable to model the temporal relationship inherent in this data as the performance of these models doesn't degrade with time as compared to the recurrent models. The mean and confidence intervals were computed over ten runs of the experiments.

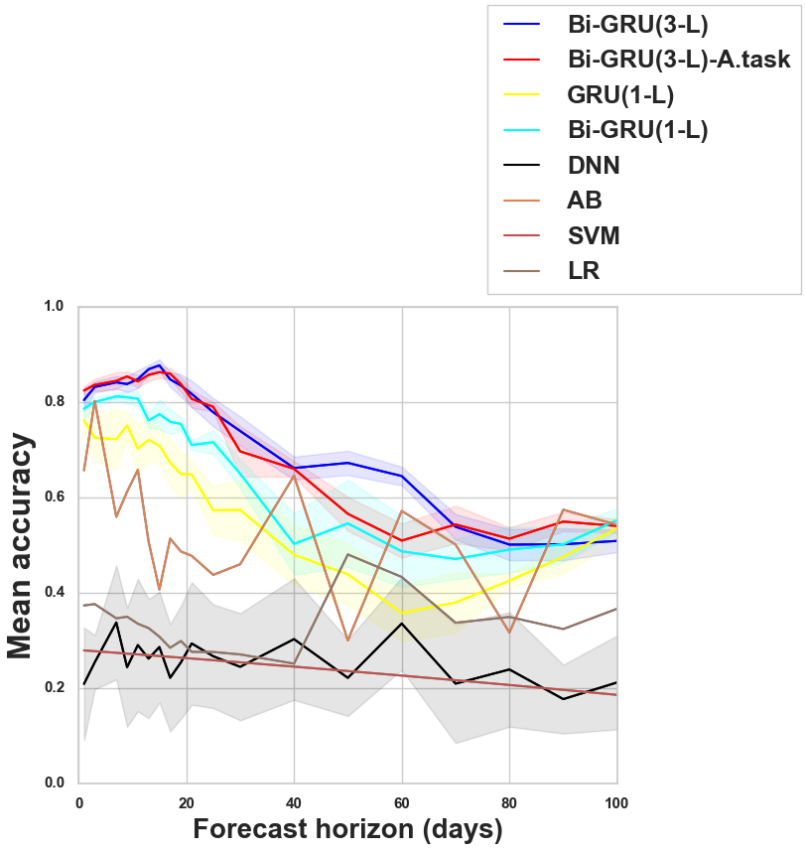


Figure 5.7: **Performance comparison (Morongo) over longer horizons.** Proposed approach with and without the auxiliary task can be seen to outperform all baselines most of the time. It can also be seen that the classical machine learning models are unable to model the temporal relationship inherent in this data as the performance of these models doesn't degrade with time as compared to the recurrent models. The mean and confidence intervals were computed over ten runs of the experiments.

	Onset of			
	Breeding	Fall Migration	Wintering	Spring Migration
1 day in advance	A,24	L,20	A,24	L,7
3 days in advance	A,46	L,20	A,12	L,35
7 days in advance	A,40	A,82	L,12	L,101

Table 5.7: **Difference between ground truth and proposed model prediction with respect to the onset of breeding, Fall Migration, wintering and Spring Migration forecasted one day, three days and seven days in advance.** All figures are in hours with A & L representing ahead and late respectively. See Figures 5.8, 5.9 and 5.10 below for extensive comparison of different sequence lengths.

	Year	Onset of Breeding	Onset of Fall migration	Onset of Wintering	Onset of Spring migration
Train	2006/2007	11/04, 03:00	28/09, 21:00	15/10, 04:00	30/03, 20:00
	2007/2008	12/04, 02:00	01/10, 01:00	18/10, 20:00	26/03, 16:00
	2008/2009	12/04, 21:00	08/10, 19:00	26/10, 00:00	06/03, 02:00
Test	2009/2010	30/03, 02:00	03/10, 18:00	25/10, 00:00	11/03, 18:00

Table 5.8: **Date (dd/mm) and time of all migration states across three years in the training and test data (Rosalie).** It can be seen that migration started and ended around a window which seems to be consistent on the average across the two continents. All times are in UTC.

	Year	Onset of Breeding	Onset of Fall migration	Onset of Wintering	Onset of Spring migration
Train	2006/2007	10/04, 04:00	07/10, 20:00	12/11, 19:00	—
Train	2007/2008	20/04, 04:00	13/09, 17:00	11/02, 18:00	22/03, 15:00
Train & Test	2008/2009	11/04, 03:00	25/09, 19:00	01/03, 23:00	10/03, 22:00
Test	2009/2010	05/04, 23:00	-	-	-

Table 5.9: **Date (dd/mm) and time of all migration states across one and a half years in the training and test data (Morongo).** It can be seen that migration started and ended sometimes late to very late during the day to very early hours of the morning. The onset of spring migration in 2006/2007, as well as the Onset of fall and spring migration as well as wintering in 2009/2010 were omitted as there were missing data points around this period. All times are in UTC.

	Year	Onset of Breeding	Onset of Fall migration	Onset of Wintering	Onset of Spring migration
Train	2007/2008	-	27/09, 20:00	18/01, 22:00	17/03, 15:00
Test	2008/2009	18/05, 22:00	13/09, 17:00	-	-

Table 5.10: **Date (dd/mm) and time of all migration states across one and a half years in the training and test data (Mac).** It can be seen that migration started and ended sometimes late to very late during the day. All times are in UTC.

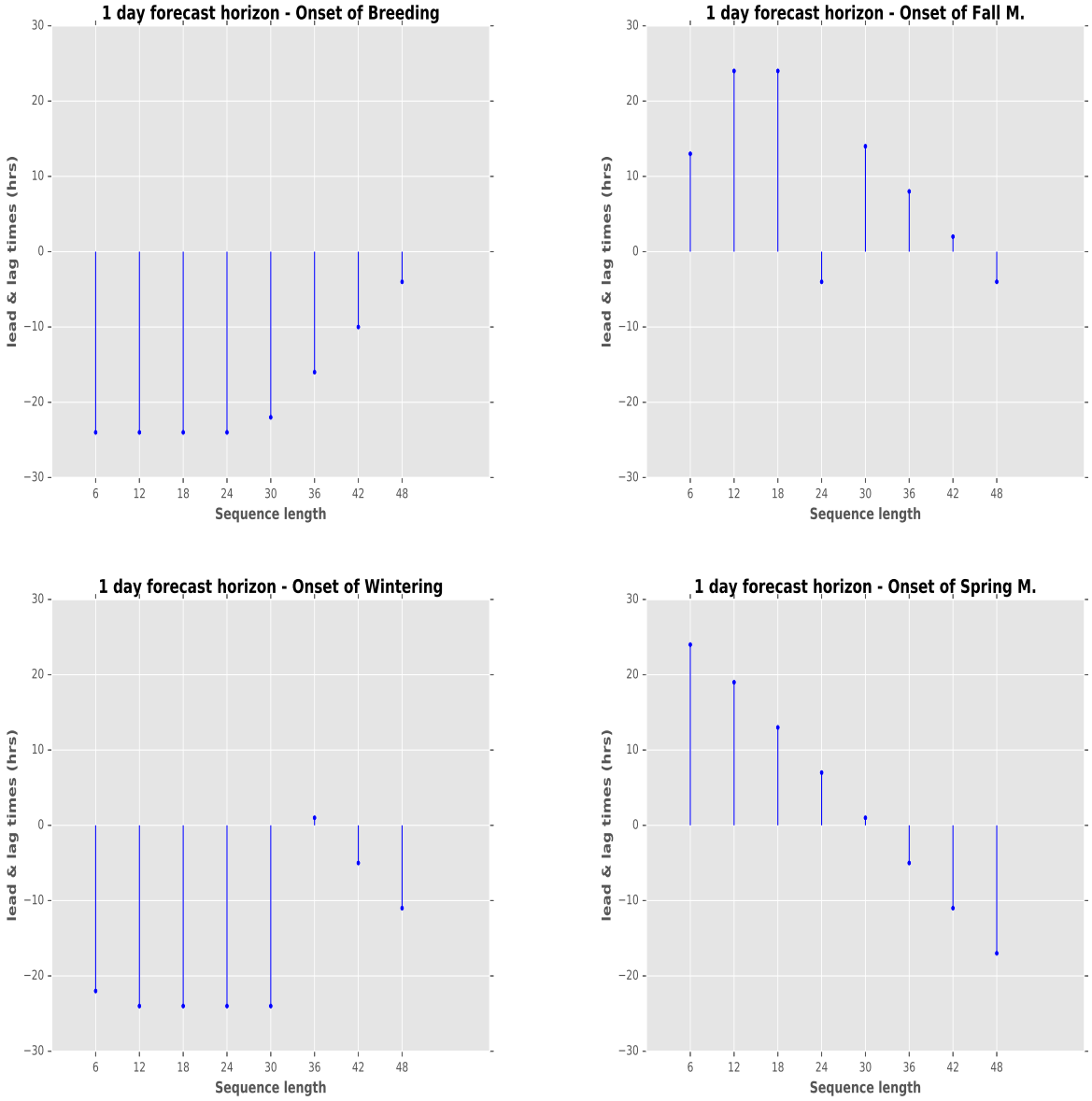


Figure 5.8: **The difference in the estimated onset of the different migration states with respect to the ground truth evaluated using several sequence lengths in Algorithm 2 for forecast horizon of 1 day.** It can be seen that the forecast horizon of 1 day gives good results when the sequence length is higher on the average.

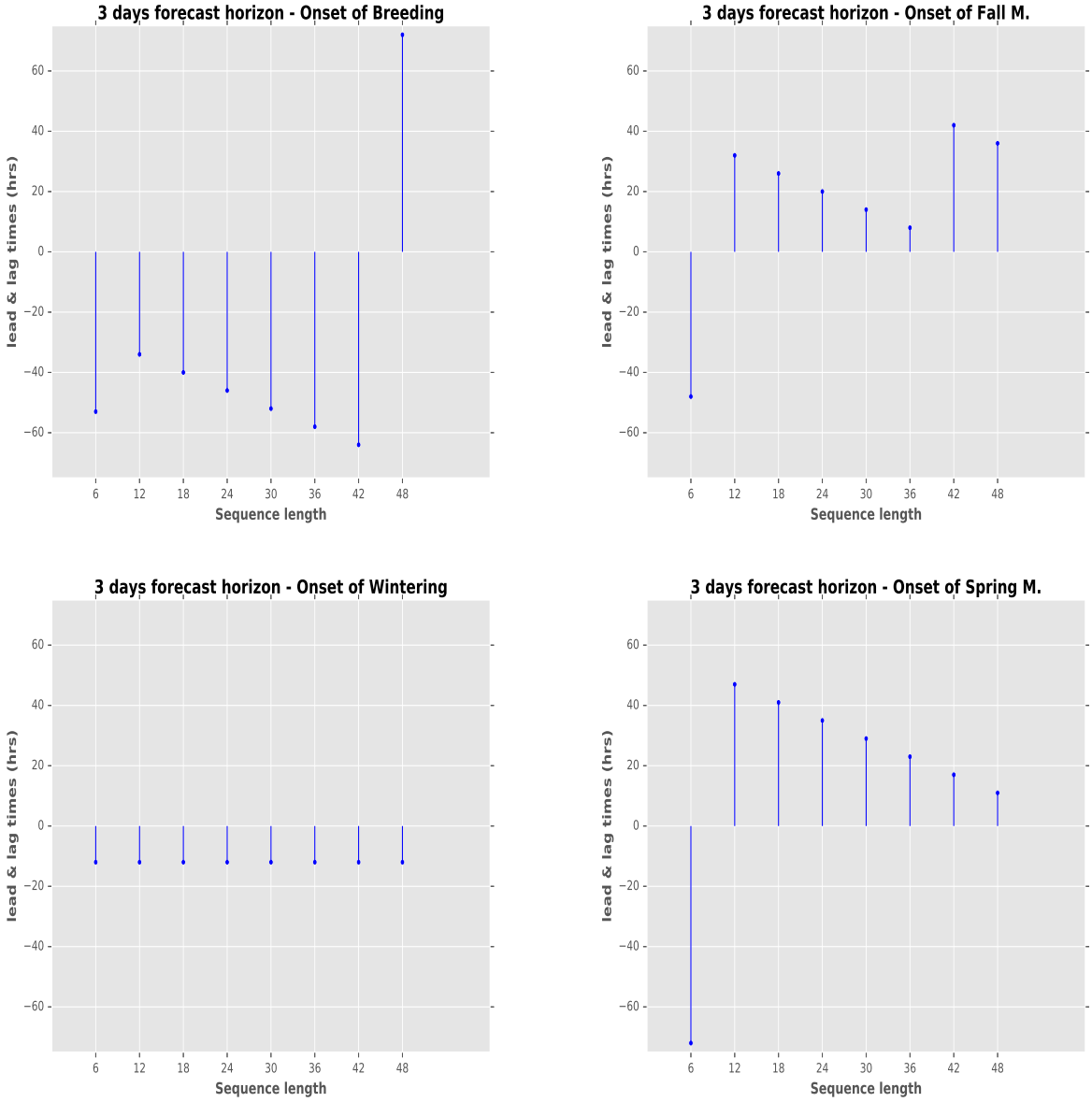


Figure 5.9: **The difference in the estimated onset of the different migration states with respect to the ground truth evaluated using several sequence lengths in Algorithm 2 for forecast horizon of 3 days.** There is no pattern however with the sequence lengths when forecasting three days in advance.

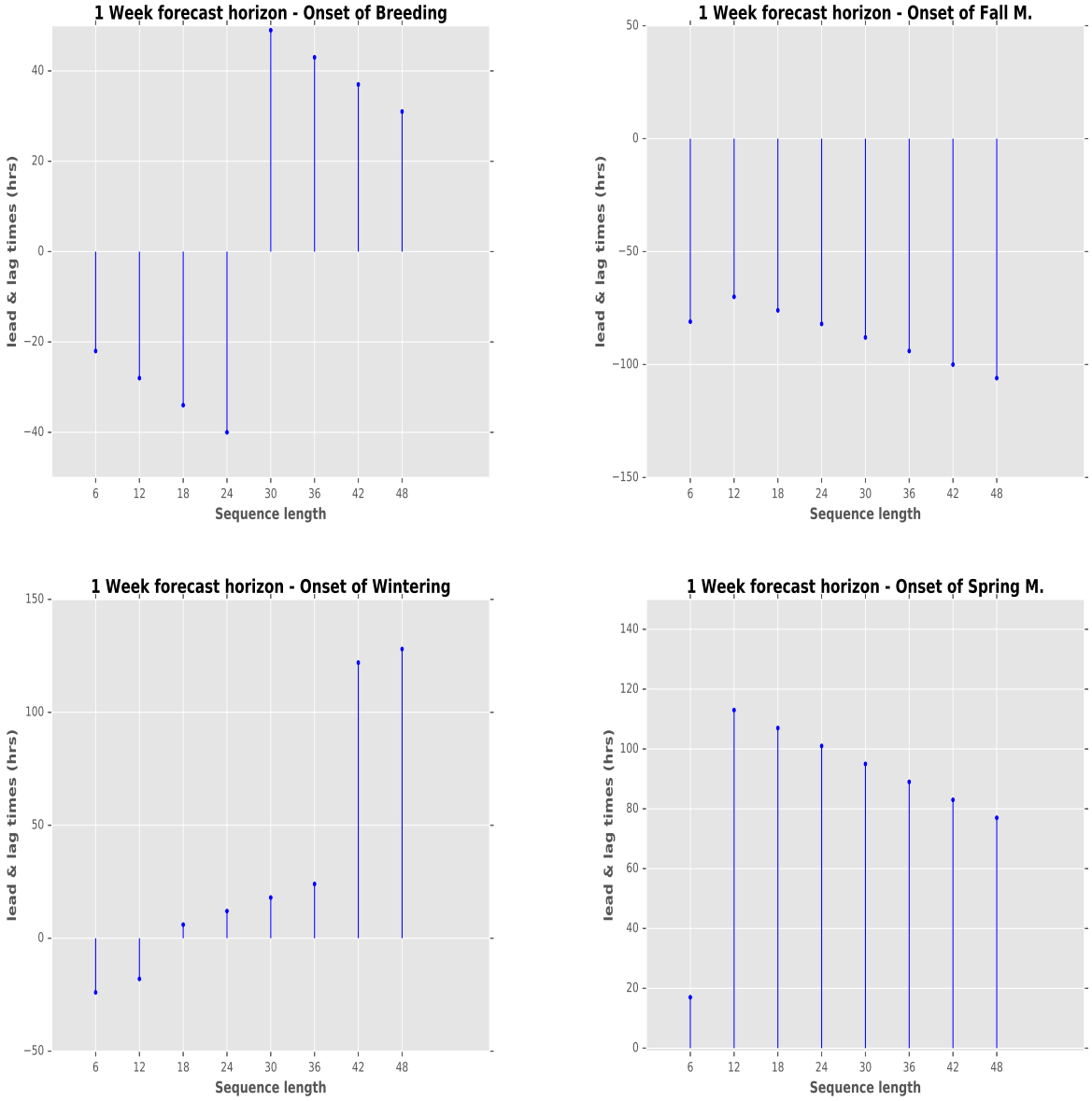


Figure 5.10: **The difference in the estimated onset of the different migration states with respect to the ground truth evaluated using several sequence lengths in Algorithm 2 for forecast horizon of 7 days.** It can be seen that small sequence length gives better results.

5.7 Limitations & Conclusion

I have proposed a deep Bidirectional-RNN deep learning network augmented with an auxiliary task towards forecasting migration patterns in Turkey vulture. This approach yield accuracies better than a variety of baselines on a variety of experiments most of the time. Ablation experiment also show this approach perform better than its variant without auxiliary task most especially over longer horizons. The implication of this work transcend reducing avian biodiversity loss. As the largest number of bird strikes occur during spring and fall migration, optimally scheduling flights during this period is a benefit this work offers which can help prevent accident and save human lives. While the prediction of fine grained avian migration patterns is a challenging task due to variability in environmental factors as a result of habitat loss, lack of access to endogenous factors coupled with the limited size of dataset[12] that I have used, the approach proposed here may be useful towards reducing avian biodiversity loss by optimally managing the infrastructures leading to these deaths within a reasonable time window. Investigating in more detail in the future the contribution of each environmental factor and additional ones will be key towards accounting for the uncertainty produced by climate change and human development. The thresholding algorithm though model specific, can also benefit from more fine-tuning and I leave that as a future work.

In this chapter, I have demonstrated the first use of supervised methods in this thesis towards forecasting migration states in migratory birds. In the next chapter, I will show the second use of supervised methods in this thesis towards forecasting stop-over decisions in migratory birds.

[12]The number of birds considered also small because of this reason.

Chapter 6

Forecasting Stop-over Decisions of Migratory Birds

In the previous chapter, I demonstrated the first use of supervised methods in forecasting avian migration patterns in this thesis. In this chapter, I address the third question raised in this thesis: Can I build better models to forecast the stop-over decisions of migratory birds at a finer granularity across different continents? Here, I show the second application of supervised methods in this thesis towards forecasting stop-over decisions of migrating birds and, by extension, also map out the stop-over sites using pre-trained supervised models to understand their characteristics.

6.1 Introduction

The one health initiative of the World Health Organization aims to control diseases (zoonotic infections) at the interface of human, environmental and animal interactions using a multidisciplinary approach, with the recent event affecting global health bringing to the fore again the increasing interdependency in the complex relationships between human and animals. This relationship is complicated by the encroachment into animal habitats by humans due to developmental activities and climate change. Major human infectious diseases such as Ebola and avian flu have been shown to have animal origin, with 20% of this coming from primates alone [231]. The spread of zoonotic infections, for instance the highly pathogenic H5N1 avian influenza,

has economic and health implications most importantly for the developing countries of the world. In light of recent developments, it is therefore becoming increasingly imperative that relevant disease surveillance and public health authorities make adequate preparations towards reducing future global health risks using holistic intervention schemes incorporating current and potential relationships between all players in nature's ecosystem.

A large number of migratory bird species stop-over intermittently to refuel, rest and re-energize [232] for the rest of the long journey as they seek warmer climes for survival during winter and spring migration (also known as flight for survival), although some of these stopovers have been attributed to deliberate plans to travel in line with the availability of resources [233]. The wild variants of these birds such as Geese, Swans, Waterfowls, Eastern wood pewee, Acadian flycatcher, Yellow-green vireo and eagles however are known to be carriers of harmful bacteria, pathogens and viruses from one place to another [234]. These harmful diseases are spread mostly at stop-over sites when these birds interact with resources in the ecosystem such as fresh water via droppings as well as interact with local birds and other non wild bird using the stop-over sites.

Adequate mapping [235, 236] of sites and forecasting of stop-over decisions have potential benefits and could help relevant national and global health authorities come up with several intervention schemes in their response and preparedness measures [237].

While previous works have largely either focused on using historical stop-over sites as a yardstick for predicting future ones [70] using Markov chain, or estimate important environmental factors relevant for stopover decisions [71], I point out here that such methods become inadequate given sites that have never been seen before in the training data with the likelihood becoming even higher given climate change activities. The approach considered in this work instead integrates this information whilst being flexible enough to potentially predict sites that have never been used before. Also, a direct benefit of this approach is the ability to forecast the duration of the stop-over. Furthermore, I aim to map out the characteristics and features of these sites towards understanding while they are chosen for stop-over. Apart from revealing potential stop-over sites in the event of extreme climate change affecting the decision to use historical sites, this can be useful for instance in creating artificial ones in nearby or isolated environments for reasons of sustainability and disease containment.

In this chapter, the task of forecasting the stop-over decisions of migratory birds for the purpose of disease outbreak management, prevention and containment is considered. I cast the task as a binary classification task where I aim to forecast whether an animal is going to move or otherwise. Given the instances where

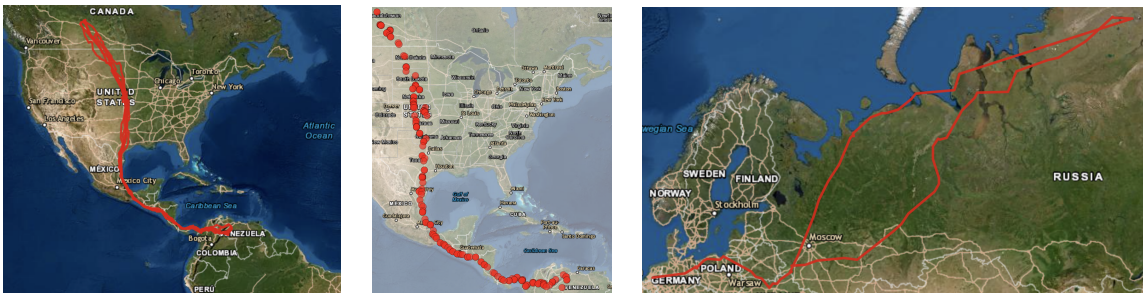


Figure 6.1: **Flyways.** Left: Trajectory of the Turkey vulture as it moves from its breeding site in north America to its wintering site in South America. Middle: Waypoints (stopover) sites of the Turkey vulture. Right: Trajectory of a White Fronted Goose as it moves between its wintering site in the Netherlands to its breeding site in the Russian Arctic.

the animals do move are in the minority in the dataset, I propose a pseudo loss objective function where the original binary cross entropy loss is recast as a multi-class cross entropy loss. I use this approach with the aid of a deep bidirectional recurrent neural network and show that it outperform several baselines most times on a variety of experiments. In addition, I also show using a publicly available convolutional neural network model pre-trained on several land cover images, characteristic of these stop-over sites and provide explanations as to what makes these sites desirable and different from non stop-over sites. All methods are evaluated on real world mobility datasets of migratory birds, the former on fine-grained (temporal) Turkey vulture mobility dataset and the latter on both the Turkey vulture dataset and that of White Fronted Geese (*Anser albifrons*) [238, 239].

The rest of the work is organized as follows: first, I provide a description of the proposed approach and datasets. This is followed by the experiments as well as a result and discussion section and I conclude by highlighting some limitations of this work and potential future directions.

6.2 Model

Given a time series $\{X_{\leq t}, y_{mt}\}_{t=n}^{T}$ where $X_{\leq t} \in \mathbb{R}^d$ and y_{mt} is a multi-dimensional vector representing input features $(X_{\leq t} = X_{t-d}, ..., X_t)$[1] and discrete stop-over decision states respectively at each time-step t where $m = 2$, the goal is to estimate $\{y_{m(t+k)}\}_{t=n}^{T}$ given $\{X_{\leq t}\}_{t=n}^{T-k}$ where k is the forecasting horizon.

[1] d is the duration of the temporal context relevant for the prediction task.

6.2.1 Deep Bidirectional RNN

This is the same model described in the previous chapter and I discuss it again here. The Bi-directional recurrent neural network [224] learns both backwards and forward context for improved sequence modelling compared to conventional recurrent neural networks. For increased expressiveness, several layers of this network can be stacked resulting in deep architectures. Given that the GRU equations in [223] can be implemented by the function G, the deep Bi-GRU with input $\mathbf{x}$, n layers and forward $(\overrightarrow{h_t})$ sequence, backward $(\overleftarrow{h_t})$ sequence as well as output y_t can then be updated as follows:

$$\begin{aligned} \overrightarrow{h_t}^n &= G(\overrightarrow{h}^n_{t-1}, \overrightarrow{h}^{n-1}_t) \\ \overleftarrow{h_t}^n &= G(\overleftarrow{h}^n_{t+1}, \overleftarrow{h}^{n-1}_t) \\ y_t &= W_{\overrightarrow{h}}\overrightarrow{h_t}^n + W_{\overleftarrow{h}}\overleftarrow{h_t}^n + b \\ h_t^{n-1} &= \mathbf{x}\ when\ n = 1 \end{aligned} \tag{6.1}$$

6.2.2 Pseudo Labels, Losses & Training Objectives

A number of phenomena in nature are imbalanced. For instance, the tendency of a natural disaster to happen, firing rates and spike bursts in the brain all have imbalanced distributions. Using this raw distribution with statistical learning models most times would result in bias outcomes and has generated huge research interest in recent times [240]. One clever approach towards reducing the effect of this imbalance without destroying temporal dependencies inherent in the data is to create pseudo labels for the majority class by dividing this class into multiple instances along dimensions that are unique within the class in relation to the input features, whilst being close to the distribution of the minority class as much as possible, but discriminative enough in this regards with respect to the input features. The binary cross entropy loss objective is given by:

$$-\left(y\log(p) + (1-y)\log(1-p)\right) \tag{6.2}$$

Where y represents the class and p the observation probability. Instead of minimizing the binary cross entropy objective loss function, given input X at training time, I aim to find θ that minimizes the overall training objective which is essentially a multi-class cross entropy loss objective within a binary cross entropy

loss objective.

$$\mathcal{L}_\theta = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{k}(y_{m,i}|x_{\leq j})\log\left(\frac{p_m}{\beta_i}\right) - \frac{1}{n}\sum_{j=1}^{n}(1-y_m|x_{\leq j})\log\left(1-p_m\right) + \lambda\mathcal{R}(\theta_e) \tag{6.3}$$

where $y_{m,i}$ represents the false labels/classes generated from the majority class y_m conditioned on the sequence of input features $x_{\leq j}$ at each data point and the associated probability distribution p_m is divided into smaller segments based on some criterion using β_i (see Algorithm 3 below) which is a function of the length of $y_{m,i}$ while $\mathcal{R}(\theta_e)$ is a regularizer with tuning hyperparameter λ. And the interpretation during training and most especially during testing is such that:

$$\begin{cases} y_m, & \text{if } \operatorname{argmax}_C \ Pr(C|X_{\leq t}) \in y_{m,i} \\ 1-y_m, & \text{otherwise} \end{cases} \tag{6.4}$$

Where C represents the new class labels after incorporating pseudo labels such that $y_{m,i} \subset C$.

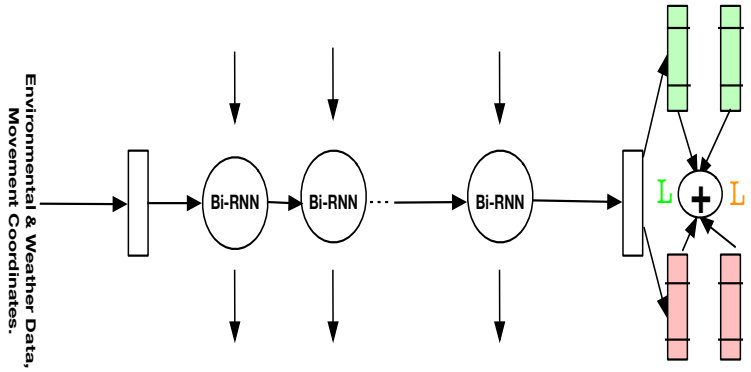


Figure 6.2: **Schematic of the proposed model.** The white rectangles represent fully connected layers while the green & red ones represent respectively the model predictions & targets for the majority class and the pseudo labels. The losses $\mathcal{L}$ are then summed up and back-propagated to learn the weights of the network.

6.3 Datasets & Preprocessing

I evaluated proposed method on datasets of Turkey vulture described in detail below and also describe the datasets of the White-Fronted Geese.

6.3.1 Turkey vulture & The White-Fronted Geese (*Anser albifrons*)

[2]This species has been identified as the vector for the avian pox [241]. Turkey vulture movement datasets [220] and [221] as part of an ongoing study were collected with the aid of GPS satellite transmitters. I obtained the trajectories of three of the birds with the fewest missing data and with sampling at a resolution of one hour as in the previous chapter. These birds breed in Washington and Oregon in the United States as well as Saskatchewan in Canada. The diversity of the breeding sites considered provides an opportunity to investigate the performance of the proposed approach and the baselines along different dimensions most especially as regards the migration route.

The White-Fronted Geese on the other hand is a species of the Geese family belonging to a larger group of birds called the waterfowl. It breeds in areas with water such as ponds, rivers and wet areas and eats food such as grasses and berries. While the population of this species appears stable globally, population of other species such as *Anser erythropus* is under threat [242]. Generally, there are populations of Geese in Europe, North America (most especially in Canada) and Asia. They fly in V shaped formation likely as a strategy to save energy [243]. A Dataset of 65 greater white-fronted geese [239] migrating between western Europe and the Russian Arctic was collected over a period of 6 years. The bird species have been identified as potential vector of infectious diseases [244] and to be precise a vector of the highly pathogenic Avian influenza Virus.

6.4 Environmental, Weather & Satellite Data

6.4.1 Environmental & Weather Data

As in the previous chapter, the movement trajectories of the Turkey vulture[3] were annotated with weather and environmental data using the Env-DATA annotation service [222]. I obtained the following features: sunshine duration, snow temperature, population density, atmospheric water, albedo, downward ultraviolet radiation at the surface, plant canopy surface water at surface, incident solar radiation, elevation, dew point temperature, snow evaporation, soil water content, water vapour concentration, snow albedo, temperature parameter in canopy conductance at surface, surface solar radiation downwards, soil temperature, surface solar radiation, maximum temperature, evaporation, ice temperature, land surface temperature night, land surface temperature day, surface thermal radiation and ten metre wind gust.

[2]See previous chapter for more information about the Turkey vulture.
[3]Three of them with the least amount of missing data.

6.4.2 Satellite Images

Satellite image patches from the Sentinel-2 satellite (part of the Copernicus sentinel missions) were collected using its API. Coordinates corresponding to the first and last seen stop-over points as well as the dates were used as input. I retrieved only the most recent image patches available online relative to the last seen date of the birds at each stopover site each season. Also, I only retrieved image patches with cloud cover of not more than 5 (see section 6.5.2, Q8 below for information on how the images were used).

6.4.3 Data Preprocessing

6.4.3.1 Environmental & Weather Data

All missing data in the environmental & weather data were replaced with the last observed points in the sequence. I divided these data into two sets: 75% for training and the rest for testing (Tables 6.1, 6.2 and 6.3). The distance covered was computed using the Haversine formula. For the purpose of this task, I considered the movement state as not moving when distance between two contiguous points is less than or equal to 1km and moving if otherwise.

$$distance(km) = 2rsin^{-1} * \sqrt{\sin^2\left(\frac{la_2 - la_1}{2}\right) + \cos(la_1)\cos(la_2)\sin^2\left(\frac{lo_2 - lo_1}{2}\right)} \tag{6.5}$$

Where r is the radius of earth, equal to 6371km and la_2, la_1 represents contiguous latitude points while lo_2, lo_1 represents contiguous longitude points.

6.4.3.2 Images

The downloaded image patches were atmospherically corrected. The sentinel-2 images have 13 spectral bands spread across three resolutions of 10m, 20m and 60m. The images were downsampled (10m band:120x120 pixels, 20m band: 60x60 pixels and 60m band: 20x20 pixels) as in [245] and converted from JPEG to TIF. I removed one of each duplicated images corresponding to instances where the bird left and returned to the same stopover site.

Data	Move	N. Moving
Training	0.22	0.78
Test	0.23	0.77

Table 6.1: **Distribution of the two classes of interest (Rosalie).** Showing the decision to move class is in the minority.

Data	Move	N. Moving
Training	0.21	0.79
Test	0.15	0.85

Table 6.2: **Distribution of the two classes of interest (Mac).** Showing the decision to move class is in the minority for the dataset.

Data	Move	N. Moving
Training	0.20	0.8
Test	0.26	0.74

Table 6.3: **Distribution of the two classes of interest (Morongo).** Showing the decision to move class is in the minority.

6.5 Experiments & Procedures

I evaluated all models using accuracy and F1 scores. All deep learning models were trained three times and evaluated on the test set for the same amount of time. I reported the mean and standard deviations of these results. To recast the problem as a multi-classification task, for simplicity, I divided the majority class along the dimension of migration or movement state (see Algorithm 3 below). Here, there are three movement states corresponding to breeding, wintering and migration where the fall and spring migration states have been merged together.

Algorithm 3 Pseudo labels

Input: (y_{mt}, M^{it}), **Output:** Pseudo labels, y_{mt}
for each t in y_{mt}:
  if M^{it} = Spring or Fall M. **&** $y_{mt\{t\}} \in$ Majority class:
    $y_{mt\{t\}} \leftarrow M^{t(i_{=Fall\ M.})}$;
  elif $M^{it} \neq$ Spring or Fall M. **&** $y_{mt\{t\}} \in$ Majority class:
    $y_{mt\{t\}} \leftarrow M^{t(i_{=Breeding,\ Wintering})}$;
  else:
    $y_{mt\{t\}} \leftarrow$ Minority Class
  end
end
return y_{mt}

6.5.1 Model Architecture & Parameters

I used a four layer[4] bidirectional recurrent neural network scanning in both directions implemented using the gated recurrent unit [223] where the input representing the data described in the environmental & weather data section above including the movement coordinates were fed to the network using a many to one

[4]This produced better results compared to smaller number of layers.

architecture. The LSTM [111] can also be used here as the performance of both when compared were the same. The hyper-parameters used include the Adam optimizer [230], cell units = 60, time step = 60, batch size = 100 learning rate = 1e-2 and a softmax at the output layer. No significance tests were carried out.

6.5.2 Experiments

I designed a number of experiments described in detail below to answer several questions that can help shed light with respect to the objectives of this work[5]. The experiments were divided broadly into two so as to assess the performance of the proposed model and as well answer some important biological and ecological questions.

- **Q1:** How better can I forecast the stopover decisions of this bird using proposed approach with respect to some baselines? To do this, I used the proposed approach and compared the performance with several baselines described below using the Turkey vulture datasets.
- **Q2:** What is the influence of the depth of the model/network on performance? Using the Turkey vulture datasets, I compared the proposed approach with a 1-layer Bidirectional recurrent network as an ablation study.
- **Q3:** How does the proposed approach perform when used without the false labels? I used this as an ablation and compared the proposed model with one trained using a binary cross entropy loss using the Turkey vulture datasets.
- **Q4:** How do the baseline models perform even when used with pseudo labels? I investigated if the proposed loss objectives can improve the performance of the baseline models using the Turkey vulture datasets.
- **Q5:** How do other techniques for handling unbalanced data perform on sequential data compared to the proposed approach? Using the Turkey vulture movement datasets, I compared techniques such as under-sampling with the proposed approach. For instance I set class weight to balanced for random forest and logistic regression and use weighted cross entropy loss for the deep learning models .
- **Q6:** Can I estimate the duration of stopover of the Turkey vulture bird and how does this compare with the ground truth? To answer this question, I compared the results of the model relative to the

[5] see Section 6.1

ground truth at each stop-over period greater than or equal to 2 hours relative to the ground truth data. Only the bird (Rosalie) with the biggest trajectory size and minimum missing data was used here to minimize uncertainty as much as possible.

- **Q7:** While the nature of the breeding and wintering habitats of the Turkey vulture is known, I ask, what are the peculiar features of the Turkey Vulture stop-over sites? To answer this question, I extracted instances where there have been significant time spent during stop-overs in the dataset. See Q8 below for more detailed description of the process involved. Only the bird (Rosalie) with the biggest trajectory size and minimum missing data was used here for brevity[6].

- **Q8:** What are the features peculiar to the stop-over sites of the White Fronted Geese? I used the pre-trained ResNet-50 architecture with a fully connected layer at the output consisting of 19 neurons (model of [245]). The labels/classes were provided from the CORINE Land Cover database of the year 2018 (CLC 2018) and merged into 19 groups for easy classification[7]. Since this is a multi label classification task, I selected a probability threshold ≥ 0.1 to observe what land cover/habitats dominates in the image patches. I validated the output of this model by cross-checking it with some of the coordinates that are well known locations/places online (wikipedia for instance). Only the bird with the biggest trajectory size in the flock per year was considered here.

6.5.3 Baselines

I compared the proposed approach to several baselines (considered to be competitive) including deep and classical learning methods[8]. In addition, a number of ablation experiments were carried out.

- **Markov Chain (MC):** Markov chain is a probabilistic model for a sequence of states where the probability of one state is dependent on the previous state. I used Markov chain as used in [70] to forecast stop-over sites but this time around to forecast stop-over decisions with the transition probabilities designed around the two mobility states.

- **Deep Neural Network (DNN):** A feed-forward network with two layers and 256 neurons each, dropout = 0.2, batch size = 100, softmax layer at the output and Adam opimizer [230].

- **Recurrent Neural Network (RNN):** A GRU network with 1 layer, 60 cells, dropout = 0.2, softmax

[6] I leave the rest as future work as it takes a lot of time to download the satellite images.
[7] The satellite images downloaded in this work were fed as input to the model so that that the model can predict its labels.
[8] Grid search used to estimate some of the parameters.

layer at the output and Adam optimizer [230]. I trained the neural networks given the procedures described for the proposed approach above.

- **Bidirectional-RNN (GRU):** As the first ablation baseline, I used a deep bidirectional recurrent neural network with just one layer and all other parameters described above for the proposed approach and without the pseudo loss objective.
- **Logistic Regression (LR):** This model uses a logistic function to model a binary dependent variable by predicting the odds of each class in the variable conditioned on the independent input variables.
- **Support Vector Machine (SVM):** This algorithm categorizes data according to optimal hyperplanes that maximizes the distance separating the classes of interest given the training data. I used a one versus one strategy for SVM.
- **Random Forest (RF):** The Random Forest algorithm [246] is an ensemble of individual decision trees where each tree predicts the class when used in a classification framework and the class with the most votes becomes the model's prediction. I also used random forest as it was used in [71] to determine what environmental factors are important for stop-over decisions with maximum depth set to 2.

No of hours in advance	Performance Metric	Models							
		LR	RF	SVM	DNN	MC	RNN (1-L)	Bi-RNN (1-L)	Proposed Model
1	Accuracy (%)	77.86	78.57	76.77	76.77	73.63	76.74	76.75	**81.76±0.123**
	Accuracy - Migration Window (%)	75.22	74.44	68.11	68.11	74.81±0.85	67.57	67.57	**86.74±0.47**
	F1 scores	(0.87,0.36)	(0.87,0.23)	(0.87,0)	(0.87,0)	(0.83, 0.43)	(0.87,0)	(0.87,0)	**(0.88, 0.59±0.01)**
	F1 scores- Migration Window	(0.84, 0.41)	0.84, 0.37)	(0.81, 0)	(0.81, 0)	(0.82, 0.57±0.01)	(0.8, 0)	(0.8, 0)	**(0.91, 0.79)**
6	Accuracy (%)	76.72	76.76	76.76	76.77	64.89	76.74	76.75	**78.19±0.25**
	Accuracy - Migration Window (%)	68.63	68.96	68.96	68.96	60.64±0.64	67.53	67.53	**79.67±0.2**
	F1 scores	(0.87,0)	(0.87, 0)	(0.87,0)	(0.87,0)	(0.77,0.23)	(0.87,0)	(0.87,0)	(0.87, **0.36±0.01**)
	F1 scores- Migration Window	(0.81, 0)	(0.82, 0)	(0.82, 0)	(0.82, 0)	(0.73, 0.26±0.02)	(0.8, 0)	(0.8, 0)	**(0.86, 0.64)**
18	Accuracy (%)	76.94	76.76	76.76	76.76	64.83	76.73	76.76	**77.44±0.392**
	Accuracy - Migration Window (%)	70.76	69.31	69.31	69.31	60.47±0.48	67.42	67.42	**75.07**
	F1 scores	(0.87,0)	(0.87, 0)	(0.87,0)	(0.87,0)	(0.77, 0.22)	(0.87,0)	(0.87,0.1)	(0.87, **0.29**)
	F1 scores- Migration Window	(0.82, 0)	(0.82, 0)	(0.82, 0)	(0.82, 0)	(0.73, 0.27±0.01)	(0.8, 0)	(0.8, 0)	**(0.83, 0.55)**
24	Accuracy (%)	76.52	76.75	76.75	76.75	64.81±0.5	76.73	76.73	76.95±0.55
	Accuracy - Migration Window (%)	68.94	68.82	68.82	68.82	60.36±0.6	67.35	67.35	**73.37±0.18**
	F1 scores	(0.86,0)	(0.87, 0)	(0.87,0)	(0.87,0)	(0.77, **0.22**)	(0.87, 0)	(0.87, 0)	(0.87, 0.17)
	F1 scores- Migration Window	(0.81, 0.15)	(0.81, 0)	(0.81, 0)	(0.81, 0)	(0.73, 0.28±0.02)	(0.8, 0)	(0.8, 0)	**(0.82, 0.47±0.04)**

Table 6.4: **Performance comparison (Rosalie).** The proposed approach can be seen to outperform all baselines across the four experiments. The F1 scores from left to right is for the decision not to move and otherwise. The figures in bold correspond to the best performing model. Confidence intervals omitted when less than 0.01. Bold numbers correspond to the best performing model.

6.6 Results & Discussion

I discuss the results of the experiments as well as the answers to the questions asked above here in the order listed.

No of hours in advance	Performance Metric	Models							
		LR	RF	SVM	DNN	MC	RNN (1-L)	Bi-RNN (1-L)	Proposed Model
1	Accuracy (%)	74.19	74.38	74.38	74.38	77±0.3	74.35	74.35	**83.55±1.2**
	Accuracy-Migration window only(%)	71.20	70.77	70.77	70.77	77.40±0.58	70.71	70.71	**84.84±0.4**
	F1 scores	(0.85,0)	(0.85,0)	(0.85,0)	(0.85,0)	(0.85,0.53)	(0.85,0)	(0.85,0)	**(0.89±0.01, 0.67)**
	F1 scores- Migration window	(0.83,0)	(0.83,0)	(0.83,0)	(0.83,0)	(0.84, 0.58±0.01)	(0.83,0)	(0.83,0)	**(0.89, 0.72±0.02)**
6	Accuracy (%)	74.38	74.38	74.38	74.38	64.76±0.43	74.34	74.16	**78.83**
	Accuracy -Migration window only(%)	70.80	70.80	70.80	70.80	62.30±0.39	70.71	70.41±0.27	**79.37±1.4**
	F1 scores	(0.85,0)	(0.85,0)	(0.85,0)	(0.85,0)	(0.77,0.22)	(0.85,0)	(0.85,0)	(**0.86±0.02**, **0.54±0.06**)
	F1 scores- Migration window	(0.83,0)	(0.83,0)	(0.83,0)	(0.83,0)	(0.75, 0.23)	(0.83,0)	(0.82, 0.15±0.06)	**(0.86, 0.59±0.01)**
18	Accuracy (%)	74.27	74.44	74.44	73.51	65.07±0.33	74.33	74.33	**75.36**
	Accuracy - migration window only(%)	70.66	70.83	70.83	70.83	62.66±0.3	70.71	70.71	71.69±1.39
	F1 scores	(0.85,0)	(0.85,0)	(0.85,0)	(0.85,0)	(0.85, **0.22**)	(0.85,0)	(0.85,0)	(0.85,0.18±0.04)
	F1 scores- Migration window	(0.83,0)	(0.83,0)	(0.83,0)	(0.83,0)	(0.75, **0.23**)	(0.83,0)	(0.83,0)	(0.83, 0.13±0.09)
24	Accuracy (%)	74.32	74.40	74.40	74.40	65±0.62	74.32	74.32	74.56±0.25
	Accuracy - Migration window only(%)	70.77	70.88	70.88	70.88	63.04±0.29	70.71	70.71	71.75±0.9
	F1 scores	(0.85,0)	(0.85,0)	(0.85,0)	(0.85,0)	(0.77, **0.22**)	(0.85,0)	(0.85,0)	(0.85, 0.1±0.05)
	F1 scores - Migration window	(0.83,0)	(0.83,0)	(0.83,0)	(0.83,0)	(0.75, 0.24±0.01)	(0.83,0)	(0.83,0)	(0.83, 0.22±0.13)

Table 6.5: **Performance comparison (Morongo).** Proposed approach can be seen to perform better than the baselines most time across the four experiments. The F1 scores from left to right is for the decision not to move and otherwise. The figures in bold correspond to the best performing model. Confidence intervals omitted when less than 0.01. Bold numbers correspond to the best performing model.

No of hours in advance	Performance Metric	Models							
		LR	RF	SVM	DNN	MC	RNN (1-L)	Bi-RNN (1-L)	Proposed Model
1	Accuracy (%)	85.24	85.24	85.24	85.24	83.28± 0.44	85.17	85.17	**88.58**
	Accuracy - Migration window (%)	80.86	80.66	80.66	80.66	81.44±0.8	80.62	80.62	**88.34±0.23**
	F1 scores	(0.92,0)	(0.92,0)	(0.92,0)	(0.92,0)	(0.9, 0.47±0.02)	(0.92,0)	(0.92,0)	**(0.93,0.61±0.02)**
	F1 scores- Migration window (%)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.88, 0.53±0.01)	(0.89, 0)	0.89, 0)	(0.93, **0.68±0.04**)
6	Accuracy (%)	85.24	85.24	85.24	85.24	70± 0.24	85.17	85.17	**86.06**
	Accuracy - Migration window (%)	80.72	80.72	80.72	80.72	67.58±0.6	80.62	80.62	82.20±1.74
	F1 scores	(0.92,0)	(0.92, 0)	(0.92,0)	(0.92,0)	(0.82,0.17± 0.01)	(0.92,0)	(0.92,0)	(0.92, **0.55±0.04**)
	F1 scores- Migration window (%)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.8, 0.21)	0.89, 0)	0.89, 0)	(0.9, 0.39±0.19)
18	Accuracy (%)	85.19	85.35	85.22	85.22	69.81± 0.27	85.17	85.17	**85.93**
	Accuracy - Migration window (%)	80.53	81.04	80.59	80.59	67.41±0.4	80.62	80.62	81.71±1.3
	F1 scores	(0.92,0)	(0.92, 0)	(0.92,0)	(092,0)	(0.81, 0.18)	(0.92,0)	(0.92,0)	(0.92,**0.35±0.09**)
	F1 scores- Migration window (%)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.79, 0.21)	0.89, 0)	0.89, 0)	(0.89±0.01, **0.51±0.05**)
24	Accuracy (%)	85.21	85.21	85.21	85.21	70.34± 0.57	85.13	85.13	**85.48**
	Accuracy - Migration window (%)	80.52	80.52	80.52	80.52	67.13±0.8	80.62	80.62	81.63±1.3
	F1 scores	(0.92,0)	(0.92, 0)	(0.92,0)	(092,0)	(0.82, 0.19±0.01)	(0.92,0)	(0.92,0)	(0.92,**0.31±0.11**)
	F1 scores- Migration window (%)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.79, 0.21)	0.89, 0)	0.89, 0)	(0.89, **0.35±0.14**)

Table 6.6: **Performance comparison (Mac).** Proposed approach can be seen to outperform all baselines across the four experiments. The F1 scores from left to right is for the decision not to move and otherwise. The figures in bold correspond to the best performing model. Confidence intervals omitted when less than 0.01. Bold numbers correspond to the best performing model.

No of hours in advance	Performance Metric	Models						
		LR	RF	SVM	DNN	MC	RNN (1-L)	Bi-RNN (1-L)
1	Accuracy (%)	77.23	77.65	76.75	76.77±0.38	73.62±0.47	75.41±1.18	75.26±1.15
	Accuracy - Migration Window (%)	76.55	82.22	67.89	67.88	74.11±0.98	59.55±1.4	63.50±4.58
	F1 scores	(0.86,0.4)	(0.87,0.19)	(0.87,0)	(0.87,0)	(0.83,0,42±0.01)	(0.86±0.01, 0.05±0.06)	(0.85±0.01, 0.15±0.1)
	F1 scores - Migration Window (%)	(0.85, 0.47)	(0.88, 0.73)	(0.81, 0)	(0.81, 0)	(0.87, 0.58±0.02)	(0.73±0.02, 0.14±0.08)	(0.74±0.05 0.36±0.07)
6	Accuracy (%)	75.17	76.46	76.76	76.77	64.35±0.89	75.32±0.89	73.24±0.83
	Accuracy - Migration Window (%)	62.18	66.96	67.74	67.73	33.44 ±1.43	62.78±2.44	61.76±1.94
	F1 scores	(0.86, 0.07)	(0.87, 0.02)	(0.87, 0)	(0.87, 0)	(0.77,0.22±0.01)	(0.86, 0.11±0.03)	(0.84, 0.27±0.05)
	F1 scores - Migration Window (%)	(0.75, 0.2)	(0.8, 0)	(0.81, 0)	(0.81, 0)	(0.49±0.01, 0.24±0.02)	(0.75±0.03, 0.25±0.13)	(0.72±0.03, 0.39±0.06)
18	Accuracy (%)	76.17	76.76	76.76	76.76	65±0.47	75.63±0.87	73.79±0.14
	Accuracy - Migration Window (%)	68.19	65.62	65.62	65.62	19.03±0.9	63.42±1.3	68.02±2.7
	F1 scores	(0.86,0.20)	(0.87, 0)	(0.87, 0)	(0.87, 0)	(0.77, 0.22)	(0.86±0.01, 0.06±0.05)	(0.84,0.26±0.03)
	F1 scores - Migration Window (%)	(0.8, 0.2)	(0.79, 0)	(0.79, 0)	(0.79, 0)	(0.29, 0.24±0.02)	(0.76, 0.14±0.03)	(0.71±0.06, 0.42±0.06)
24	Accuracy (%)	75.79	76.68	76.75	76.74	64.51±0.29	74.06±1.41	71.12±1.5
	Accuracy - Migration Window (%)	67.26	64.02	64.58	64.58	16.27±0.6	60.41±2.4	60.52±2.5
	F1 scores	(0.86,0.21)	(0.87, 0)	(0.87, 0)	(0.87, 0)	(0.77, 0.22)	(0.85±0.01, 0.06)	(0.82±0.01, 0.19±0.05)
	F1 scores - Migration Window (%)	(0.8, 0.31)	(0.78, 0)	(0.78, 0)	(0.78, 0)	(0.23, 0.26)	(0.74±0.02, 0.1±0.04)	(0.73±0.03, 0.26±0.06)

Table 6.7: **Performance comparison of baselines with pseudo labels (Rosalie).** Overall, the performance of the baseline models can be seen on the average to have degraded slightly when used with pseudo labels.

No of hours in advance	Performance Metric	Models						
		LR	RF	SVM	DNN	MC	RNN (1-L)	Bi-RNN (1-L)
1	Accuracy(%)	73.71	74.75	74.38	74.38	77.26	74.31	74.46
	Accuracy-migration window only(%)	71.13	70.71	70.71	70.71	75.20±0.36	70.58±0.19	70.51±0.16
	F1 scores	(0.84,0.24)	(0.85,0)	(0.85,0)	(0.85,0)	(0.85,0,53)	(0.85, 0)	(0.85, 0)
	F1 scores - Migration window	(0.82, 0.23)	(0.83, 0)	(0.83, 0)	(0.83, 0)	(0.86, 0.57)	(0.83, 0)	(0.82, 0)
6	Accuracy (%)	71.16	74.44	74.38	74.38	64.70	74.26	72.28±1.09
	Accuracy-migration window only(%)	66.86	70.55	70.55	70.55	43.27±0.7	70.71	70.86±0.71
	F1 scores	(0.82,0.25)	(0.85,0	(0.85,0)	(0.85,0)	(0.77,0.22)	(0.85, 0)	(0.83, 0.19±0.02)
	F1 scores - Migration window	(0.78, 0.27)	(0.83, 0)	(0.83, 0)	(0.83, 0)	(0.6, 0.24)	(0.82, 0)	(0.82, 0.12±0.01)
18	Accuracy (%)	73.28	74.40	74.38	74.38	64.66	74.35	73.10±0.9
	Accuracy-migration window only(%)	69.05	69.93	69.93	69.92	32.15±0.5	70.36±0.5	70.67±0.3
	F1 scores	(0.84,0.17)	(0.85,0)	(0.85,0)	(0.85,0)	(0.77,0.22)	(0.85,0)	(0.83±0.01, 0.23±0.1)
	F1 scores - Migration window	(0.81, 0.16)	(0.82, 0)	(0.82, 0)	(0.82, 0)	(0.49, 0.23)	(0.82, 0)	(0.81±0.02, 0.28±0.23)
24	Accuracy (%)	72.78	74.40	74.40	74.40	65.31±0.18	74.26	73.69
	Accuracy-migration window only(%)	67.86	69.74	69.74	69.74	29.46±0.12	70.48±0.2	69.99±0.5
	F1 scores	(0.84, 0.11)	(0.85, 0)	(0.85, 0)	(0.85, 0)	0.77, 0.23±0.01	(0.85, 0)	(0.84, 0.10±0.05)
	F1 scores - Migration window	(0.8, 0.12)	(0.82, 0)	(0.82, 0)	(0.82, 0)	(0.46, 0.22)	(0.82, 0)	(0.82, 0)

Table 6.8: **Performance comparison of baselines with pseudo labels (Morongo).** Overall, the performance of the baseline models can be seen on the average to have degraded slightly when used with pseudo labels. The F1 scores from left to right is for the decision not to move and otherwise.

No of hours in advance	Performance Metric	Models						
		LR	RF	SVM	DNN	MC	RNN (1-L)	Bi-RNN (1-L)
1	Accuracy (%)	82.92	85.24	85.24	85.24	83.88±0.75	83.80±0.86	85.16
	Accuracy - Migration window (%)	81.73	80.66	80.66	80.66	80.86±0.49	78.24±3.4	79.90±1.14
	F1 scores	(0.9,0.25)	(0.92,0)	(0.92,0)	(0.92,0)	(0.9, 0,49±0.01)	(0.9,0)	(0.92,0)
	F1 scores- Migration window (%)	(0.9, 0.25)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.9, 0.53)	(0.88±0.02, 0)	(0.88, 0)
6	Accuracy (%)	74.78	85.22	85.22	85.24	70.42	84.81±0.45	78.01±3.45
	Accuracy - Migration window (%)	69.73	80.72	80.72	80.72	53.41±0.35	80.02±0.7	73.91±4.05
	F1 scores	(0.85,0.28)	(0.92,0	(0.92,0)	(0.92,0)	(0.82 ,0.15)	(0.92,0.1±0.08)	(0.87±0.02 ,0.24±0.11)
	F1 scores- Migration window (%)	(0.8, 0.33)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.69, 0.2)	(0.89, 0)	(0.83±0.04, 0.34±0.1)
18	Accuracy (%)	79.30	85.22	85.22	85.22	70.37±0.94	85.01	79.89±3.9
	Accuracy - Migration window (%)	77.69	80.59	80.59	80.59	38.15±0.5	79.77±0.7	74.86±2.3
	F1 scores	(0.88,0.20)	(0.92,0)	(0.92,0)	(0.92,0)	(0.82, 0.17±0.01)	(0.92,0)	(0.88±0.03,0.24±0.17)
	F1 scores- Migration window (%)	(0.87, 0.21)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.55, 0.2)	(0.89, 0)	(0.85±0.02, 0.45±0.03)
24	Accuracy (%)	82.05	85.21	85.21	85.21	70±0.14	84.90±0.39	83.19±1.66
	Accuracy - Migration window (%)	79.67	80.52	80.52	80.52	36.05±1.38	80.40±0.1	80.08±0.4
	F1 scores	(0.9,0.19)	(0.92, 0)	(0.92,0)	(0.92,0)	0.82, 0.18	(0.92, 0)	(0.91±0.01, 0.1±0.03)
	F1 scores- Migration window (%)	(0.88, 0.15)	(0.89, 0)	(0.89, 0)	(0.89, 0)	(0.53±0.01, 0.19±0.01)	(0.89, 0)	(0.89, 0.12±0.07)

Table 6.9: **Performance comparison of baselines with Pseudo labels (Mac).** Overall, the performance of the baseline models can be seen on the average to have degraded slightly when used with pseudo labels. Confidence intervals are omitted when less than 0.01.

hrs in advance		1	6	18	24
Without	Accuracy (%)	81.59±0.395	78±0.2633	77.03±0.3351	76.94 ± 0.04
	Accuracy- Migration Window (%)	87.68±1.57	78.58±0.5	70.97±1.61	70.39±1.67
	F1 minority class	0.55±0.06	0.37±0.06	0.26±0.126	0
	F1 minority class - Migration Window (%)	0.79±0.04	0.61±0.03	0.24±0.11	0.22±0.11
	F1 majority class	0.88	0.87	0.86	**0.87**
	F1 majority class - Migration Window (%)	0.91	0.85	0.82	0.82
With	Accuracy (%)	81.76	78.19	**77.44**	76.95±0.455
	Accuracy- Migration Window (%)	86.74±0.47	79.67±0.2	**75.07**	**73.37±0.18**
	F1 minority class	0.59	0.36	0.29	**0.17**
	F1 minority class - Migration Window (%)	0.79	**0.64**	**0.55**	**0.47±0.04**
	F1 majority class	0.88	0.87	**0.87**	0.86
	F1 majority class - Migration Window (%)	0.91	**0.86**	**0.83**	0.82

Table 6.10: **Performance of model with and without pseudo loss (Rosalie).** The performance of the proposed approach can be seen to have degraded a bit without the pseudo loss. Bold numbers correspond to the best performing model.

hrs in advance		1	6	18	24
Without	Accuracy(%)	83.67±0.9	80±0.8	75±0.8	75.08 ± 0.05
	Accuracy-migration window only(%)	84.95±2.52	80.31±0.37	71.82±0.67	71.50±0.95
	(F1 minority class)	0.66±0.03	**0.6±0.01**	0.13±0.11	0.1 ± 0.08
	F1 minority class - migration window only(%)	0.71±0.08	**0.63±0.05**	0.15±0.12	0.13±0.13
	F1 majority class	0.89	0.86	0.85	0.85
	F1 majority class - migration window only(%)	**0.9±0.01**	0.86	0.83	0.83
With	Accuracy (%)	83.55±1.2	78.88±1.3	75.36	74.56±0.25
	Accuracy-migration window only(%)	84.84±0.4	79.37±1.4	71.69±1.39	**71.75±0.9**
	(F1 minority class)	**0.67**	0.54±0.06	**0.18±0.04**	0.1±0.05
	(F1 minority class - migration window only(%)	0.72±0.02	0.59±0.01	0.13±0.09	**0.22±0.13**
	F1 majority class	0.89±0.01	0.86±0.02	0.85	0.85
	F1 majority class - migration window only(%)	0.89	0.86	0.83	0.83

Table 6.11: **Performance of model with and without pseudo loss (Morongo).** The performance with and without the pseudo labels is almost the same.

hrs in advance		1	6	18	24
Without	Accuracy(%)	88	87±1.41	83.33±0.47	84.53 ± 1.11
	Accuracy- Migration window	87.57±0.4	**84.58±0.32**	82.90±3.1	82.44±0.8
	F1 minority class	0.61±0.02	**0.56±0.01**	**0.47±0.03**	0.29±0.13
	F1 minority class-Migration window	0.65±0.02	**0.6±0.04**	0.42±0.13	**0.46±0.04**
	F1 majority class	0.93	0.92	0.92	0.92±0.01
	F1 majority class-Migration window	0.92	0.9	0.89	0.82
With	Accuracy (%)	**88.58**	86.06	**85.93**	85.48
	Accuracy - Migration window (%)	**88.34±0.23**	82.20±1.74	81.71±1.3	81.63±1.3
	F1 minority class	0.61±0.02	0.55±0.04	0.35±0.09	**0.31±0.11**
	F1 minority class-Migration window	**0.68±0.04**	0.39±0.19	**0.51±0.05**	0.35±0.14
	F1 majority class	0.93	0.92	0.92	0.92
	F1 majority class-Migration window	**0.93**	0.9	0.89±0.01	**0.89**

Table 6.12: **Performance of model with and without pseudo loss (Mac).** The performance between the two models can be seen to be almost the same due to the small size of dataset used. Confidence intervals are omitted when less than 0.01.

- **A1:** Results in Tables 6.4, 6.5 and 6.6 show I can forecast the mobility decisions with the proposed approach better than a variety of baselines. Not only are the accuracies better but the F1 scores for the minority class across the experiments are better. In addition, the F1 scores for migration period only, show that the stop-over decisions using proposed model can be estimated with good scores.
- **A2:** The increase in the depth of the model/network helped improve the performance with and without the use of the pseudo loss objectives (See Tables 6.4, 6.5, 6.6, 6.7, 6.8, 6.9, 6.10, 6.11 and 6.12).
- **A3:** Results in Tables 6.10, 6.11 and 6.12 show that without the pseudo labels, the proposed approach on the average outperforms its version without the pseudo loss objective most especially over longer forecast horizon.
- **A4:** Results in Table 6.7, 6.8 and 6.9 show that the performance of the baselines degrades slightly on

No of hours in advance	Performance Metric	Models					
		LR	RF	DNN	RNN (1-L)	Bi-RNN (1-L)	Bi-RNN (4-L)
1	Acc.(%)	75.98	75.37	76.77	24.35	32.82	79± 0.01
	Acc.- Migration Window(%)	83.11	85.11	73.40±7.49	32.09±0.4	55.981.1	84.22±2.3
	F1 scores	(0.83,0.57)	(0.82,0.59)	(0.87,0)	(0, 0.372)	(0.22±0.1,0.4±0.01)	(0.85±0.01, 0.65±0.01)
	F1 scores - Migration Window	(0.88, 0.72)	(0.89, 0.77)	(0.83±0.03, 0.26)	(0, 0.48±0.01)	(0.35±0.22, 0.54±0.03)	(0.87±0.02, 0.79±0.02)
6	Acc.(%)	56.13	56.08	76.74	24.87	50.81±0.1	70± 0.02
	Acc.- Migration Window(%)	55.28	53.61	68.96	34.5±1.9	55.78±2.46	74.62±3
	F1 scores	(0.64,0.44)	(0.63,0.46)	(0.87,0))	(0,0.376)	(0.78± 0.02,0.45± 0.03)	(0.54± 0.16,0.56±0.01)
	F1 scores - Migration Window	(0.58, 0.52)	(0.51, 0.56)	(0.82, 0)	(0, 0.49)	(0.55±0.05, 0.56±0.01)	(0.78±0.03, 0.7±0.02)
18	Acc.(%)	67.57	63.70	76.75	23.59	54.15±0.1	63.77± 0.014
	Acc.- Migration Window(%)	67.41	65.96	68.27±1.47	32.92±0.4	48.77±11.75	62.88±7.51
	F1 scores	(0.77,0.44)	(0.72,0.46)	(0.87,0)	(0,0.376)	(0.6±0.11,0.44±0.02)	(0.71±0.02,0.5)
	F1 scores - Migration Window	(0.75, 0.52)	(0.73, 0.53)	(0.81±0.01, 0)	(0, 0.49)	(0.31±0.28, 0.52±0.03)	(0.64±0.1, 0.61±0.04)
24	Acc.(%)	67.22	66.64	76.75	23.34±0.1	23.51±0.27	57.6±4.2
	Acc.- Migration Window(%)	70.95	72.18	68.82	32.61±0.4	33.41±1.07	64.35±1.42
	F1 scores	(0.77, 0.42)	(0.76, 0.44)	(0.87, 0)	(0, 0.38)	(0, 0.37)	(0.65, 0.46)
	F1 scores - Migration Window	(0.78, 0.55)	(0.79, 0.57)	(0.81, 0)	(0, 0.49)	(0, 0.48±0.02)	(0.66±0.02, 0.62)

Table 6.13: **Performance comparison (Rosalie).** Baseline models when used with classical data imbalance techniques. Some of the results are not monotonic due to the destruction of the temporal dependency in the data by the use of special resampling and re-weighting techniques thus rendering these methods not so useful for tasks such as forecasting (an exception is the proposed model without the pseudo loss due to its deep architecture). Confidence intervals are omitted when less than 0.01.

No of hours in advance	Performance Metric	Models					
		LR	RF	DNN	RNN (1-L)	Bi-RNN (1-L)	Bi-RNN (4-L)
1	Acc.(%)	73	75	57.66±23	45.33±3.3	56.57±5.2	79.47± 2.56
	F1 scores	(0.82,0.44)	(0.81,0.61)	(0.65±0.28,0.1±0.1)	(0.51±0.1,0.32±0.14)	(0.64±0.08, 0.42±0.04)	(0.85±0.02, 0.69±0.02)
	Acc.- Migration Window(%)	73.29	76.59	70.79	30.56±0.9	41.91±6.07	75.41±6.2
	F1 scores - Migration Window	(0.83, 0.42)	(0.82, 0.65)	(0.83, 0)	(0, 0.45)	(0.39±0.17, 0.41±0.07)	(0.79±0.06, 0.69±0.05)
6	Acc.(%)	57	50	27	51.33±7.4	63.67±1.6	66.67± 1.88
	F1 scores	(0.63,0.48)	(0.51, 0.49)	(0,0.4))	(0.55±0.12,0.45±0.02)	(0.71± 0.02,0.53)	(0.73± 0.01,0.58±0.01)
	Acc.- Migration Window(%)	58.46	54.10	70.79	31.24±2	55.10±9.1	66.57±4.28
	F1 scores - Migration Window	(0.63, 0.53)	(0.53, 0.55)	(0.82, 0)	(0, 0.45)	(0.54±0.15, 0.55±0.04)	(0.7±0.05, 0.62±0.02)
18	Acc.(%)	65.76	64.64	39.03±9	52.33±13.57	57.67±9.4	60± 4.97
	F1 scores	(0.76,0.4)	(0.72,0.5)	(0.36 ±0.26,0.35)	(0.56±0.2,0.34±0.1)	(0.64±0.12,0.44±0.02)	(0.66±0.06,0.51±0.02)
	Acc.- Migration Window(%)	66.86	59.21	69.66±1.65	37.82±7.3	50.74±6.3	55.95
	F1 scores - Migration Window	(0.77, 0.43)	(0.65, 0.51)	(0.82±0.02, 0)	(0.27±0.21, 0.43±0.02)	(0.51±0.1, 0.49±0.02)	(0.58, 0.54)
24	Acc.(%)	66.33	63.84	32.48±6.1	30.59±4.17	37.53±5.4	0.51±0.11
	F1 scores	(0.78, 0.27)	(0.73, 0.43)	(0.26±0.21, 0.33)	(0.14±0.11,0.41)	(0.34±0.12, 0.39±0.02)	(0.53±0.16, 0.46±0.02)
	Acc.- Migration Window(%)	65.36	65.06	64.47±5.19	29.84±0.45	32.27±1.6	45.74±5.86
	F1 scores - Migration Window	(0.77, 0.27)	(0.74, 0.48)	(0.75±0.08, 0.35±0.24)	(0, 0.45)	(0.13±0.04, 0.45)	(0.41±0.11, 0.49±0.02)

Table 6.14: **Performance comparison (Morongo).** Baseline models when used with classical data imbalance techniques. Confidence intervals are omitted when less than 0.01. The temporal dependency in the data can be seen to have been destroyed using these techniques apart from the very deep model. The F1 scores from left to right is for the decision not to move and otherwise.

No of hours in advance	Performance Metric	Models					
		LR	RF	DNN	RNN (1-L)	Bi-RNN (1-L)	Bi-RNN (4-L)
1	Acc.(%)	73	66	39.33±19.6	42.97±5.7	54.67±4.1.	77.99± 0.8
	F1 scores	(0.82,0.47)	(0.76,0.44)	(0.43±0.31,0.24±0.17)	(0.53±0.08,0.25±0.02)	0.67±0.03,0.25±0.05)	(0.85, 0.55)
	Acc.- Migration Window(%)	78.98	70.38	73.62±0.03	30.99±2	42.04±10.02	78.39±0.02
	F1 scores - Migration Window	(0.86, 0.58)	(0.78, 0.53)	(0.81±0.03, 0.53±0.01)	(0.28±0.04, 0.33)	(0.46±0.16, 0.35±0.02)	(0.85±0.02, 0.63±0.02)
6	Acc.(%)	55	47	38.33±33	50±10.61	60.33±0.9	73.67± 1.88
	F1 scores	(0.65,0.36)	(0.56,0.35)	(0.31±0.43,0.17±0.12))	(0.59±0.14,0.32±0.02)	(0.71±0.01,0.38)	(0.82±0.02,0.51±0.01)
	Acc.- Migration Window(%)	58	47.52	57.95±27.48	19.98±0.3	58.44±3.6	73.44±3.3
	F1 scores - Migration Window	(0.66, 0.45)	(0.52, 0.42)	(0.58±0.41, 0.11±0.11)	(0, 0.32)	(0.66±0.04, 0.46±0.01)	(0.8±0.03, 0.58±0.03)
18	Acc.(%)	64	73	56.12±17.82	41.33±7.6	57.67±5.4	72.33± 3.2
	F1 scores	(0.75,0.34)	(0.83,0.4)	(0.67 ±0.16,0.25±0.1)	(0.48±0.11,0.3±0.01)	(0.68±0.05,0.37±0.02)	(0.81±0.03,0.48±0.01)
	Acc.- Migration Window(%)	66.74	75.02	42.51±16.34	24.46±0.02	59.44±3	72.38±1.1
	F1 scores - Migration Window	(0.77, 0.41)	(0.84, 0.45)	(0.42±0.3, 0.36±0.02)	(0.14±0.06, 0.32)	(0.68±0.03, 0.44±0.02)	(0.8±0.01, 0.56)
24	Acc.(%)	70.49	64.23	85.21	22±5.1	23.51±0.27	72.21±1.8
	F1 scores	(0.8, 0.38)	(0.74, 0.41)	(0.92, 0)	(0.16±0.11, 0.26)	(0, 0.37)	(0.81±0.01, 0.48)
	Acc.- Migration Window(%)	72.55	67.20	41±25.34	22.08	33.28±9.2	71.93±1.4
	F1 scores - Migration Window	(0.82, 0.44)	(0.76, 0.49)	(0.4±0.35, 0.31±0.1)	(0, 0.33)	(0.31±0.16, 0.34±0.02)	(0.8±0.01, 0.54±0.01)

Table 6.15: **Performance comparison (Mac).** Baseline models when used with other data imbalance techniques. Confidence intervals are omitted when less than 0.01.

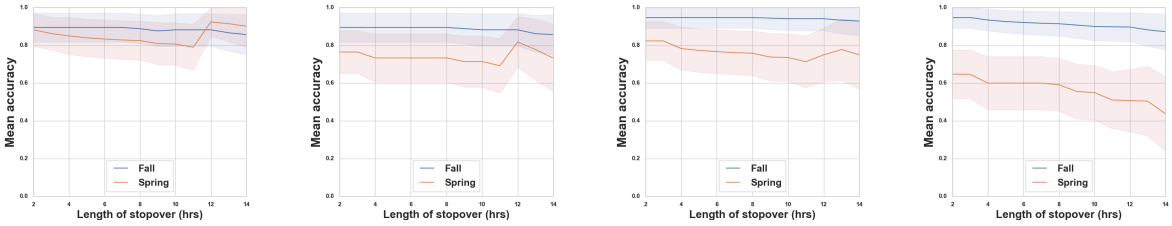


Figure 6.3: **Duration of Stopover.** Left to Right: Forecast horizon of 1, 6, 18 and 24 hours respectively. The y axis represents the mean proportion of the stopover duration on the **x**-axis that has been estimated correctly up until the first decision to move is predicted relative to the ground truth with zero being the lowest and 1 the highest. The spikes in the spring migration is related to the frequency of occurrence of the stop-over duration beyond 10 hours as the birds rarely stop for a long time during spring migration.

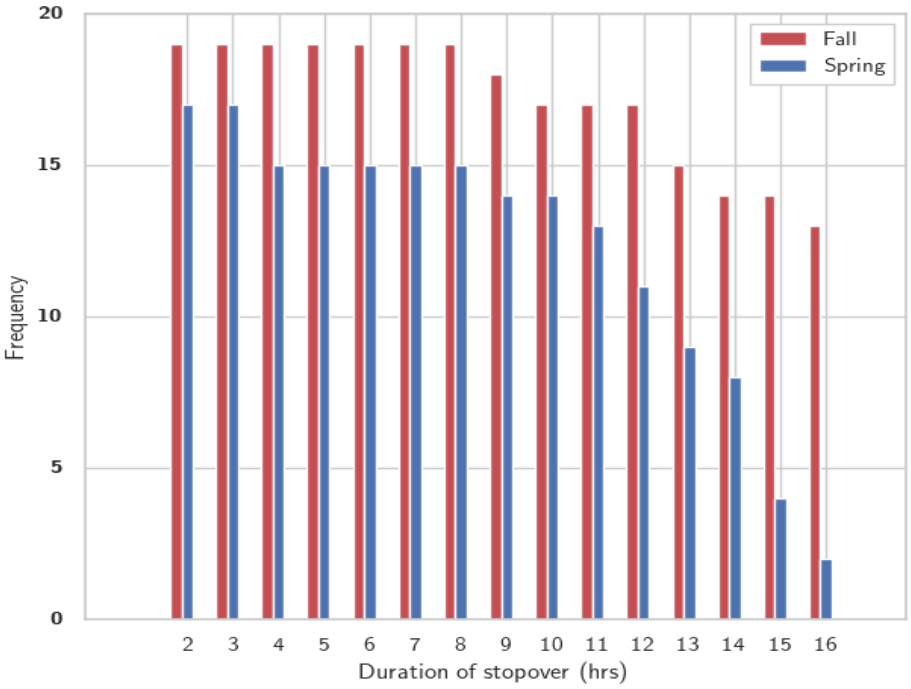


Figure 6.4: **Distribution of the stop-over duration of the Turkey vulture across the test data.** It can be seen that the frequency is higher during fall migration compared to spring. This is a common phenomena observed with other birds as well, and the reasons for this is currently an active area of research.

Figure 6.5: **Typical stopover site for Turkey vultures and White Fronted Geese**. Mixture of coniferous forest (forest with cone shaped trees) and water bodies as seen via google map. The White Fronted Geese stop at sites predominantly with access to water.

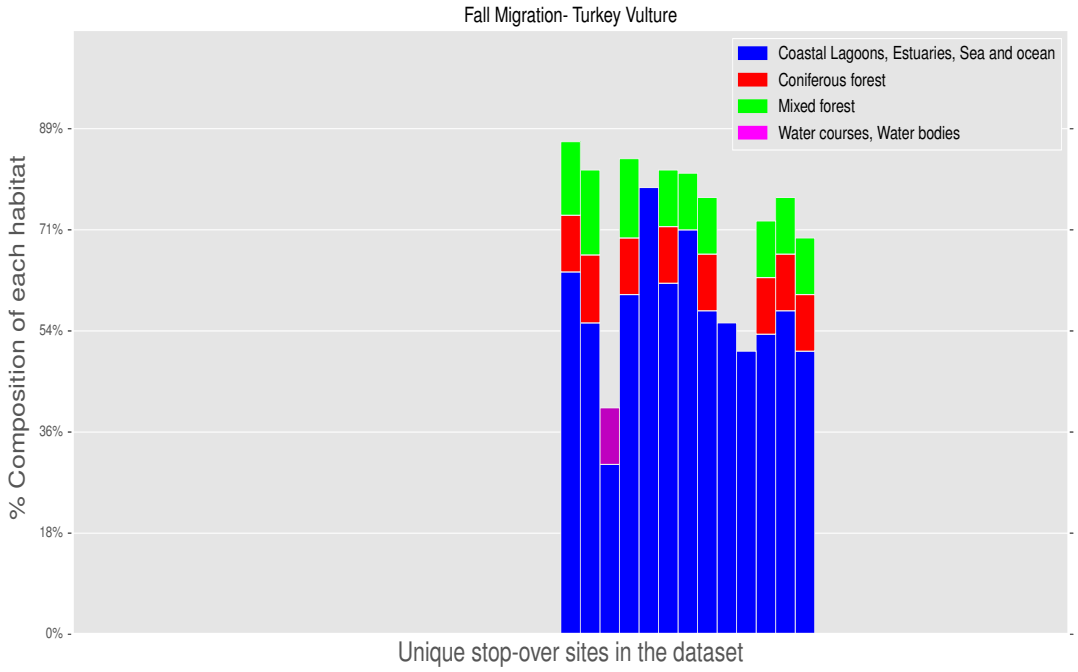
the average when used with the pseudo labels but the F1 scores for the minority class improved most especially for the deep learning sequence models. This demonstrates to some extent the generalisation of the proposed approach to other deep sequence architectures.

- **A5:** Results in Tables 6.13, 6.14 and 6.15 show that conventional techniques for handling class imbalance hurt the performance with respect to accuracy but enhanced the performance of the F1 scores for the underrepresented class at the expense of the other class for the classical models while the proposed approach was able to improve the classification of the minority class without hurting the majority. Overall, the results suggest these methods for handling class imbalance may be inappropriate for the kind of sequential data considered in this chapter.

- **A6:** Results in Figure 6.3. show the duration of stopover can be estimated with high accuracy which appears to decrease as the forecast horizon increases and on the average higher for fall migration stopovers. In addition, the accuracy is higher for the fall migration on the average because fall migration is generally longer than spring migration (see Figure 6.4) meaning more training data and more confidence in the estimation for fall relative to spring stop-overs.

- **A7:** Results show this bird stop-over at places with access to water (see Figure 6.6). While it is surprising this bird never stopped over in stereotype sites with access to carrion, Turkey vultures are known to feed on insects and washed up fish [247] and this could be a strategy to have access to both food and water simultaneously with little effort or the priority could just be on water only as some birds eat enough to sustain them before embarking on migration.

- **A8:** One particular feature common to these stop-over sites is the presence of water (see Figure 6.7). This makes sense given the affinity of this avian species to water and its surrounding areas when



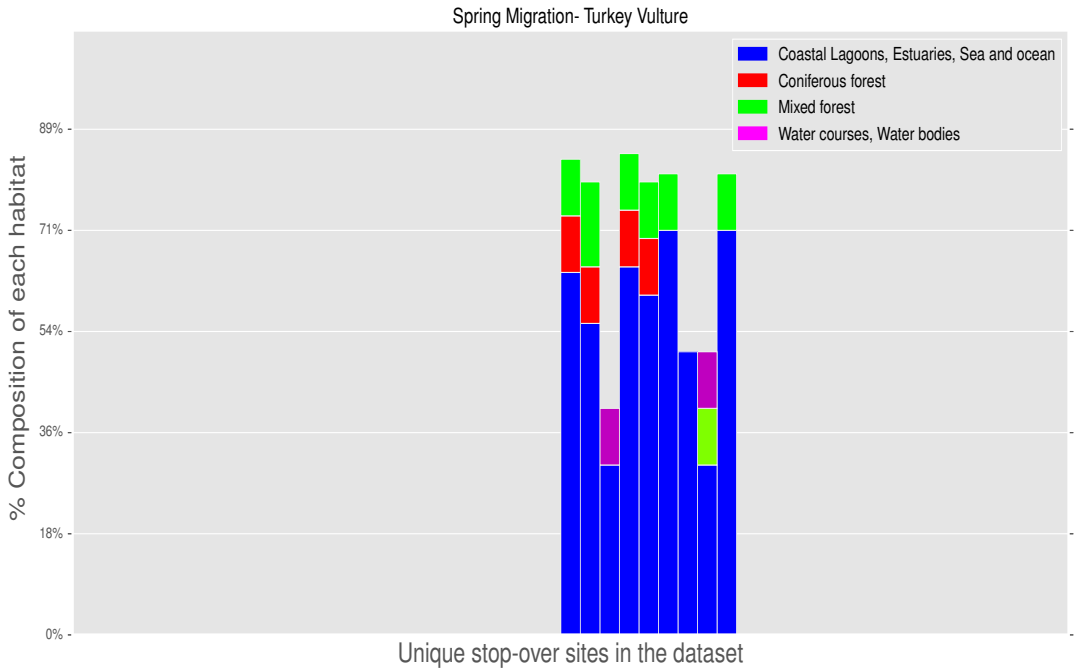


Figure 6.6: **Turkey vulture stop-over habitats.** The bird can be seen to stop over at sites with access to water bodies and resources. Each bar represents a unique stop-over period and location in the dataset where all the groups/components in each bar sum up to 100%. Only habitats with probability/composition greater than 10% are shown for brevity.

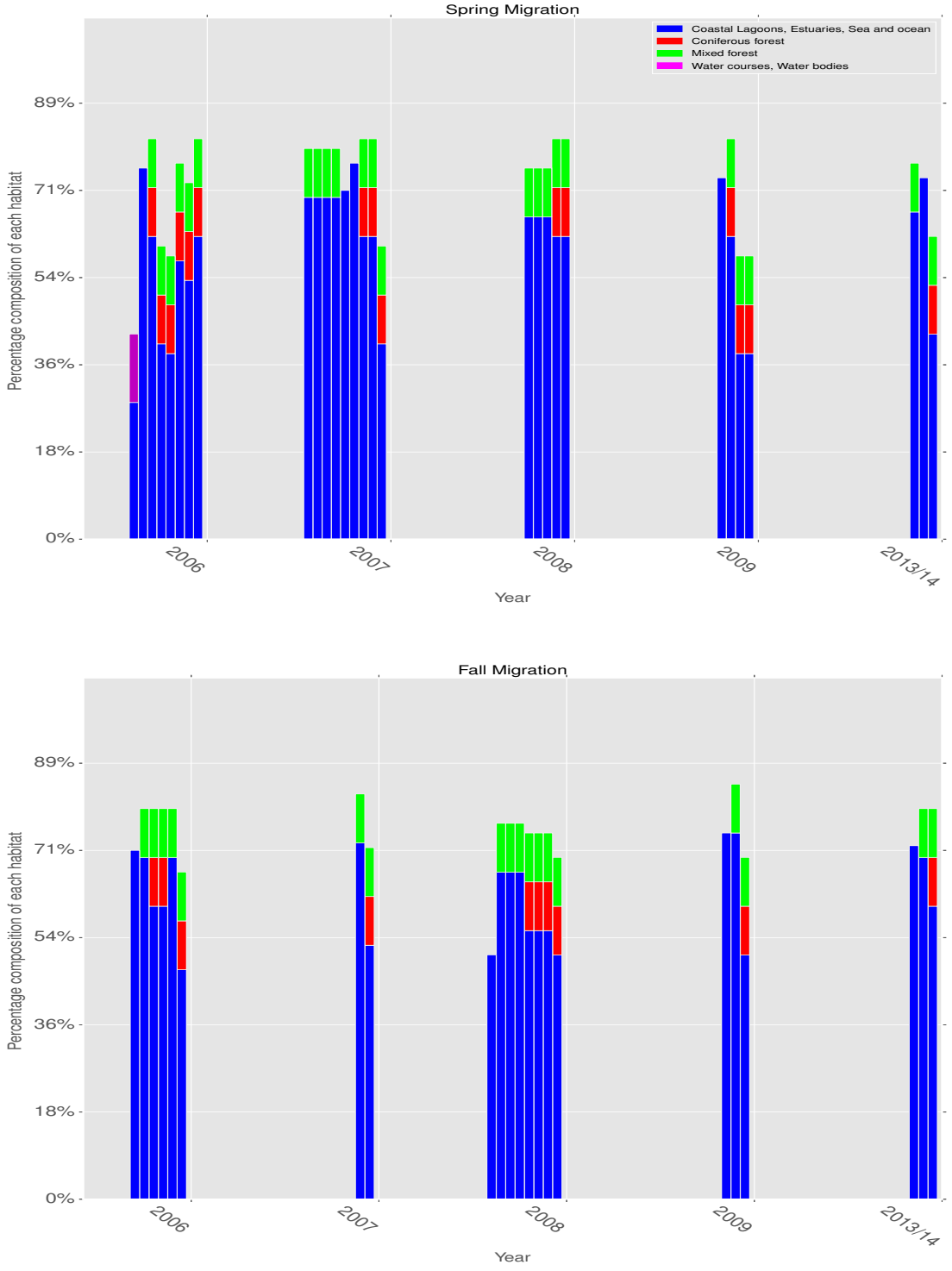


Figure 6.7: **Mapping of habitat (White Fronted Geese).** Each bar chart represents a unique stop-over period and location of the White Fronted Geese where all the groups/components in each bar sum up to 100% (Only habitats with probability/composition greater than 10% are shown for brevity).

breeding. The presence of auxiliary information about the surrounding habitat can however be helpful in what type of water body is used here. In this case, since the water bodies were mapped with coniferous and mixed forests, it is of the opinion they are not seas and oceans but smaller water bodies such as lakes and ponds as they both possess such characteristics whereas the water bodies without forest are in most cases very large rivers or seas/oceans/littoral. The risks with respect to the spread of diseases however are via droppings and interactions at water bodies in urban areas or areas with high population density most especially when such water bodies are used for domestic and recreational activities[9].

6.7 Limitations & Conclusion

The conflict between humans and animals in the usage of natural habitats and resources is on the increase with its associated health threats due to human developmental activities such as agriculture, mining, pollution via industrialization among others . In this chapter, I have proposed a bidirectional recurrent neural network with pseudo loss objectives to aid the prediction of mobility decisions of migratory birds. I also mapped out the preferred stop-over habitats using a pre-trained convolutional neural network with results showing the two species of birds considered in this work prefer stop-over sites with adequate access to water. Crucially, this work offers a new perspective to the problem of stop-over prediction in migrating birds by considering data at a finer granularity as well as understanding why the stop-over sites were chosen for this purpose. This work however is not without its limitations. First, I have ignored the prediction of movement velocities here for simplicity due to the quality and quantity of data used[10]. This will be incorporated into the model in the future as more fine-grained data collected over long period becomes available. To compensate for the lack of velocity information, I have mapped out the preferred stop-over habitat using the land cover classification paradigm. This classification model used for mapping the habitats is by no means perfect but good enough to give a high level and general description of the land cover in the image patches without much ambiguity. In addition, the forecast horizon is not practical enough (3 days to 1 week would suffix) and I aim to follow up on that in future work[11]. I also aim to look at other wild avian species using other migratory paths such as the Australasia, Central Asian and African-Eurasian flyways. Beyond global health concerns, this work has implication towards worldwide conservation efforts. Overall, I conclude by emphasizing the continuous

[9] Water bodies around bird farms should also be considered here.

[10] In practice as it stands, the approach would assume the knowledge of the stop-over habitats most especially the ones at risk of spreading diseases, with the forecasting task carried out with respect to these locations.

[11] In addition, just like in the previous chapter, the number of birds considered is small because of the size of the datasets.

global surveillance of wild migratory birds represent some of the necessary steps we must take if we must reduce some of our future global health risks.

In this chapter, I have demonstrated the second use of a supervised approach in this thesis towards forecasting the stop-over decisions of migratory birds (Turkey vulture to be precise) and also show how to predict discrete avian movement decisions. In the next chapter, I will discuss the first application of a combination of supervised and unsupervised methods using classical and modern deep learning methods towards understanding the decision making process of a sheep as well as predicting it. In addition, I will show how to combine both discrete and continuous movement decision making models in this regard.

Chapter 7

Learning to Herd

In this chapter, I address the fourth question raised in this thesis: can a hierarchical deep learning predictive model of sheep movement behaviour perform better than some alternatives? Here, I show the application of both supervised and unsupervised methods discussed in previous chapters in learning movement patterns given multimodal datasets of sheep movement behaviour over long horizons. It is hoped that by building models of real world movement behaviour of animals, their decision making process can be better understood in addition to discovering new architectures in artificial intelligence scalable into more real world problems.

7.1 Introduction & Motivation

Modern Artificial intelligence (AI) systems and algorithms have struggled on real world tasks, especially complex cognition and reasoning tasks that evolve over very long horizons. Some of the challenges facing today's AI system's include but not limited to, continual learning[1], out of distribution problems; and generalisation [248] as well as learning from small samples. The principles underlying natural intelligence which involves the intelligence and cognition in living systems such as humans and animals holds key towards solving some of the problems highlighted above. In essence, AI holds the key towards understanding natural intelligence and, at the same time, understanding natural intelligence could inform the development of new intelligent systems using AI. The benefit here is mutual.

[1] leveraging previous learning experiences to learn new tasks efficiently.

Concurrently, humans and animals develop mental models and representations of their environment based on the sensory spatio-temporal information they receive. Understanding the causal relationship between the sensory inputs animal receive and motor actions, however, has been a major goal of most behavioural scientists, psychologists, neuroscientists, ethologists and, recently artificial intelligence researchers [82]. Early work in this field was restricted to studying the behaviour of animals in well controlled and restricted laboratory environments where the natural behaviour is limited, with data analysis largely qualitative or based on human observation. Recently, due to advances in technology, low cost miniaturized sensors have become increasingly available for tracking natural organisms across several scales and, most importantly, in their natural habitats. This has made possible the study of animal behaviour in challenging environments and circumstances.

Animals visualise, smell, move, plan and coordinate differently to humans and the underpinning principles for this natural intelligence/abilities are worth investigating using a data driven approach. An understanding of these abilities is of potential importance in building systems and artificial agents with similar characteristics in the near future. Of particular interest among the behaviours exhibited and displayed by animals is the movement and coordination of sheep in a flock. Sheep are social animals with interesting movement and coordination characteristics that make them suitable for modelling decision making in intelligent animals. While the popular perception is that these creatures are not very intelligent, recent studies have shown that the true picture is more nuanced, and that they have excellent memory and recognition abilities [249]. In addition, sheep have been shown to have excellent problem solving capabilities and can recognise emotional states solely via facial characteristics [250], hence their choice for this task.

As mentioned earlier, the multimodal nature of behavioural data, however, means classical machine learning models are unable to model properly the different modalities of behaviour as these data violate the i.i.d.[2] assumption necessary for such models to perform optimally, thus creating out of distribution problems [248]. However, recent advances in deep learning have made it possible to build models that can represent more complex stochastic distributions and so can reproduce better the behaviour of intelligent animals. But, these methods still have their limitations, for instance defining the reward function in any real world task is not a trivial thing to do let alone for natural systems if a reinforcement learning approach is to be used. In addition, other methods leveraging deep learning or other classical methods have ignored the relationships between several modes of behaviour[3]. Scaling some of these methods up using ideas grounded in cognitive neuroscience represents a promising direction and is the main focus of this work. Cognition in relation to

[2]independent and identically distributed.
[3]loosely referred to here as causality

memory vis-a-vis the storage, organization and retrieval of information is key towards producing advanced intelligent behaviour, and designing architectures using cognitive principles in the context of modern deep learning systems is essential towards achieving this.

Recent works (see chapter two for more detailed review) in the area of modelling and understanding the structure of animal behaviour have already shown that animal (mouse) behaviour is made up of modules that do not exist in isolation [78] but are selected by switching the transition dynamics around these modules. The concept of modularity has also been used to model the visual intelligence of dog as it affects movement [84] as well as the movement behaviour of fruit flies [82] leveraging recent advances in neural networks. I build on the findings in these body of work in this thesis by showing how these modules can be used to generate highly complex movement behaviour in more natural environment. While simple mathematical models might suffix to model the behaviour of animals, I consider the most neuroethological(ly) plausible approach to gain a better understanding on the structure of animal movement behaviour and how information might be processed to generate these behaviours.

More specifically in this thesis, I use movement data from a flock of 36 sheep gathered with the aid of GPS devices over several days in the open. Using this data, I ask, can I build better predictive model of sheep movement abilities in a herd compared to some alternatives? To this end, first I propose a novel hierarchical architecture using recurrent neural networks for this task to process, categorize sensory inputs and map these inputs to corresponding behavioural modes. Second, I leverage recent advances in deep generative neural networks to model these sheep movement behaviour modes in a herd. More specifically, I use several recurrent neural networks and mixture density networks with specialized tasks to approximate the behaviour of a sheep where the recurrent networks help to guide the received sensory inputs to the relevant hierarchically factorized behavioural movement mode[4]. I show the generality of the proposed approach on several other datasets of the same flock collected across different locations at different times and empirically compare it to a variety of baselines. Results show that the proposed approach outperform competing baselines most of the time on a variety of evaluation metrics.

The remainder of this chapter is organized as follows. Next, I describe the overall architecture. The datasets as well as the associated preprocessing are described and this is followed by the experiments and results sections in that order. As usual, I conclude this chapter by highlighting the limitations of the methods used.

[4]Modes and modules will be used interchangeably from here.

7.2 Methods

7.2.1 Architecture

I describe the whole architecture in this section, a summary of which can be found in Figure 7.1 and is a combination of deep neural networks at different levels of hierarchy. This architecture has been inspired by previous works [82, 84] in this area where movement behaviour have been clustered using a flat approach without structures as to the relationship between clusters. Instead, this architecture uses a hierarchical clustering with recurrent networks designed to model the relationship between clusters. There is also some resemblance between this architecture and the behaviour tree [251] in that apart from their both modular structure, the internal nodes act as control flow nodes with the leaf nodes acting as execution nodes.

First, the architecture starts with a network that decides the collective behaviour of the flock as either largely stationary or largely mobile. If stationary, no further action is taken otherwise, the sensory inputs are sent to another level of decision making where the agent has to decide whether to move or not. If the agent decides not to move, no further decision is taken but if it decides to move, the agent's state is mapped to actions representative of several modalities of movement behaviour via several binary decision making networks stacked hierarchically. All the binary decision making processes were modelled using a RNN (LSTM) to add context and the velocities to be taken modelled using a MDN where a feed forward network is used to parameterise a mixture of Gaussians of the velocities given the state. I expand on some segments of the architecture below.

In the next few subsections, I will discuss some of the key components of this architecture.

7.2.1.1 Annotations

With the aid of a viewer, each time point in the trajectories of the flock were classified into two on the basis of the collective behaviour of the flock. Subjectively, it appears that the herd have two main collective behaviours. One, when the whole flock is largely stationary with little or no movement activity and a second when they have a variety of activities spread across the field. By separating the collective behaviour into two modes, the problem of modelling the movement pattern of any sheep of interest when the whole flock is largely stationary is eliminated. Each collective behaviour was assigned different labels such that a supervised model can be trained to distinguish between the two classes of herd behaviour. This was used as a prior and the performance of the model with and without this prior was investigated in the results section.

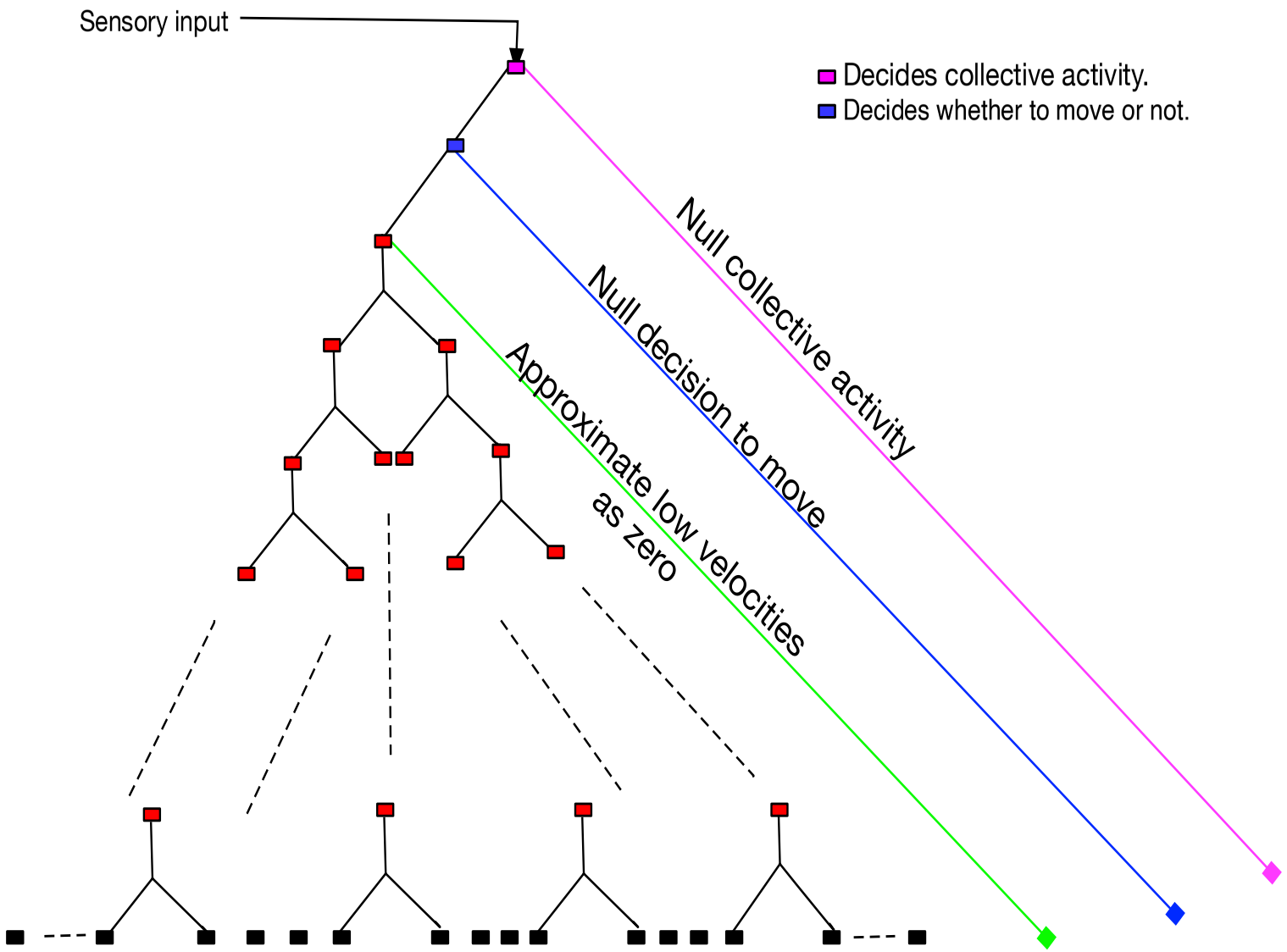


Figure 7.1: **Architecture of the hierarchical framework used.** Sensory input represents state information (see section 7.2.1.3). Each red node represents a recurrent neural network that determines the child node to select. The bottom nodes (black) are the movement modules modelled using a mixture density network.

7.2.1.2 Activity recognition

I expatiate on what I have discussed earlier on this subject. The goal here is to recognize the overall behaviour of the flock. From observations, I was able to recognize there are almost two distinct collective activities exhibited by the sheep in the training dataset and I approximated the collective behaviour along these lines. These activities were classified into when the whole herd is still with little or no movement activities within the herd and everything else apart from this to be in another class. The state/input here is the difference between the centroid of the herd relative to the positions of each sheep. I labelled the sequence of movement in the training dataset to classify the collective group activity and learn the corresponding label in the test dataset in a supervised manner using a cross entropy loss.

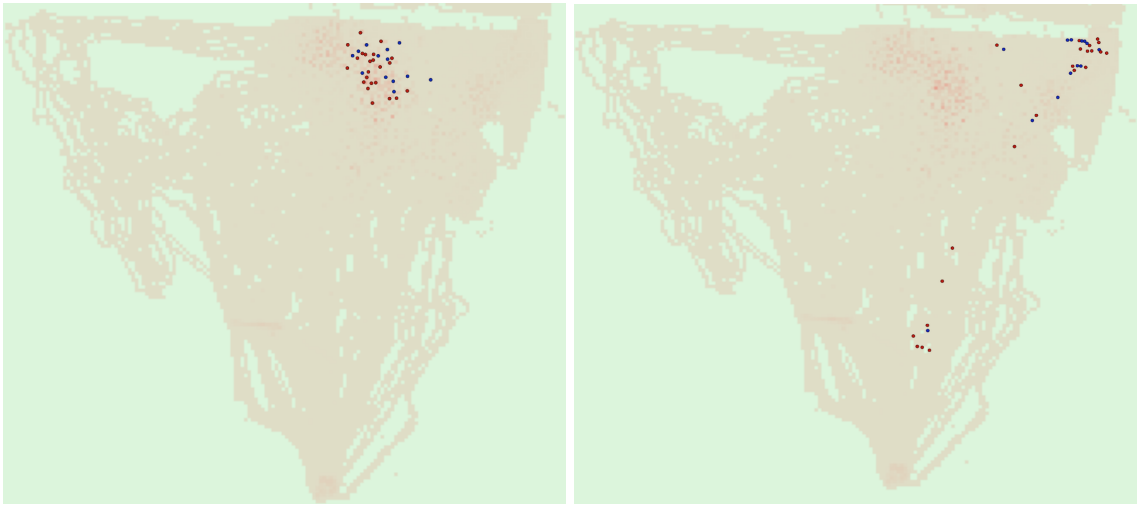


Figure 7.2: **Collective movement of the flock seen with the aid of a viewer:** Left: Sheep still with little or no movement and Right: Sheep scattered across the field with varying activities.

7.2.1.3 To move or not?

Due to the sedentary nature of sheep, it can be observed, more often than not, that when the sheep have diverse activities including feeding and moving across the field, the animals are largely stationary. Therefore, the next decision making process after deciding the state of the collective behaviour is to predict first whether an animal is going to move or not. The labels were derived from the velocities such that, if the change in x and y coordinates between two contiguous time steps were both zero, a label of zero was given to this state and one if otherwise. A recurrent neural network was trained in a supervised learning framework to map states to discrete labels. The trained supervised model for this task was further calibrated as with all model

for real world tasks. The state information[5] used here and for all subsequent tasks include:

(1) Distance from the sheep of interest to N - nearest sheep.

$$distance = \sqrt{(Y_i - Y_{j=1...N})^2 + (X_i - X_{j=1...N})^2} \tag{7.1}$$

(2) The angles between the sheep of interest and N - nearest sheep.

$$\text{atan2}(Y_i - Y_{j=1...N}, X_i - X_{j=1...N})$$

(3) Velocities of N nearest sheep to the sheep of interest with respect to the change in x and y coordinates.

$$\Delta X_{j=1...N}, \Delta Y_{j=1...N} \tag{7.2}$$

These state information were concatenated together and fed as input to all models used in this architecture apart from the GMM used for clustering. No information regarding environmental variables such as weather and atmospheric conditions were factored into the state information. I also didn't factor information about the visibility of other sheep in the state information.

7.2.1.4 To move with what velocity?

If the modelled artificial sheep decides to move, the question is with what velocity? The velocity when moving was modelled using a mixture density network where the state of the animal was used as input to a feed forward network to parameterise a Gaussian mixture of the velocities belonging to each decomposed movement module using a probabilistic framework. Here, the final decomposed behavioural/movement modules were obtained by decomposing the overall movement behaviour (velocities) of the flock hierarchically following tree patterns (see Figure 7.3). The state here is as described above while the actions are velocities derived by first sampling from a categorical distribution to determine which Gaussian component to use and the continuous value sampled from the corresponding Gaussian component.

[5]The state/sensory information has been designed to account for spatio-temporal changes in the environment of the sheep mostly as it affects the behaviour of other animals that are closest to the one of interest.

7.2.2 Overall Training Objectives

There are four key components of the proposed model overall, each with its own training objective(s). First there are three supervised learning models defined by a recurrent neural network, one for determining the collective behaviour, the second for making the binary decision to either move or not, and the third, an array of memory models arranged hierarchically, making a 2-way decision to determine what movement module to use to select the velocity given the decision to move. The last key component is the mixture density network which outputs the velocity and is trained by minimizing the negative log likelihood of the GMM likelihood function.

$$\mathcal{L}_{mdn} = \sum_{z=1}^{Z} -\log\left[\sum_{i=1}^{K} \Pi_i(x)\mathcal{N}(\mu_i(x), \sigma_i^2(x))\right]_z \tag{7.3}$$

Where Z represents the total number of movement behaviour modules used and the parameters are as previously described.

Given the cross entropy losses for selecting the collective behaviour (L_{cb}), the binary mobility decision (L_{mb}) and the movement module to activate for velocity (L_c) are given respectively by,

$$\mathcal{L}_{cb} = \sum_{n=1}^{N} L(y, f(X_n; \theta_1)) \tag{7.4}$$

$$\mathcal{L}_{mb} = \sum_{n=1}^{N} L(y, f(X_n; \theta_2)) \tag{7.5}$$

$$\mathcal{L}_{c} = \sum_{m=1}^{M}\sum_{n=1}^{N} L(y, f(X_n; \theta_{3m})) \tag{7.6}$$

Where M represents the total number of recurrent supervised models for selecting the movement module to activate and N the size of the data while X_n represents the input to the model, y the output/labels of the network while θ_1, θ_2 and θ_3 are parameters of the recurrent neural networks. The total training and loss objectives (L_{tot}) if the whole network is trained end to end is therefore a sum of the loss functions described above:

$$\mathcal{L}_{tot} = \mathcal{L}_{cb} + \mathcal{L}_{mb} + \mathcal{L}_{c} + \mathcal{L}_{mdn} \tag{7.7}$$

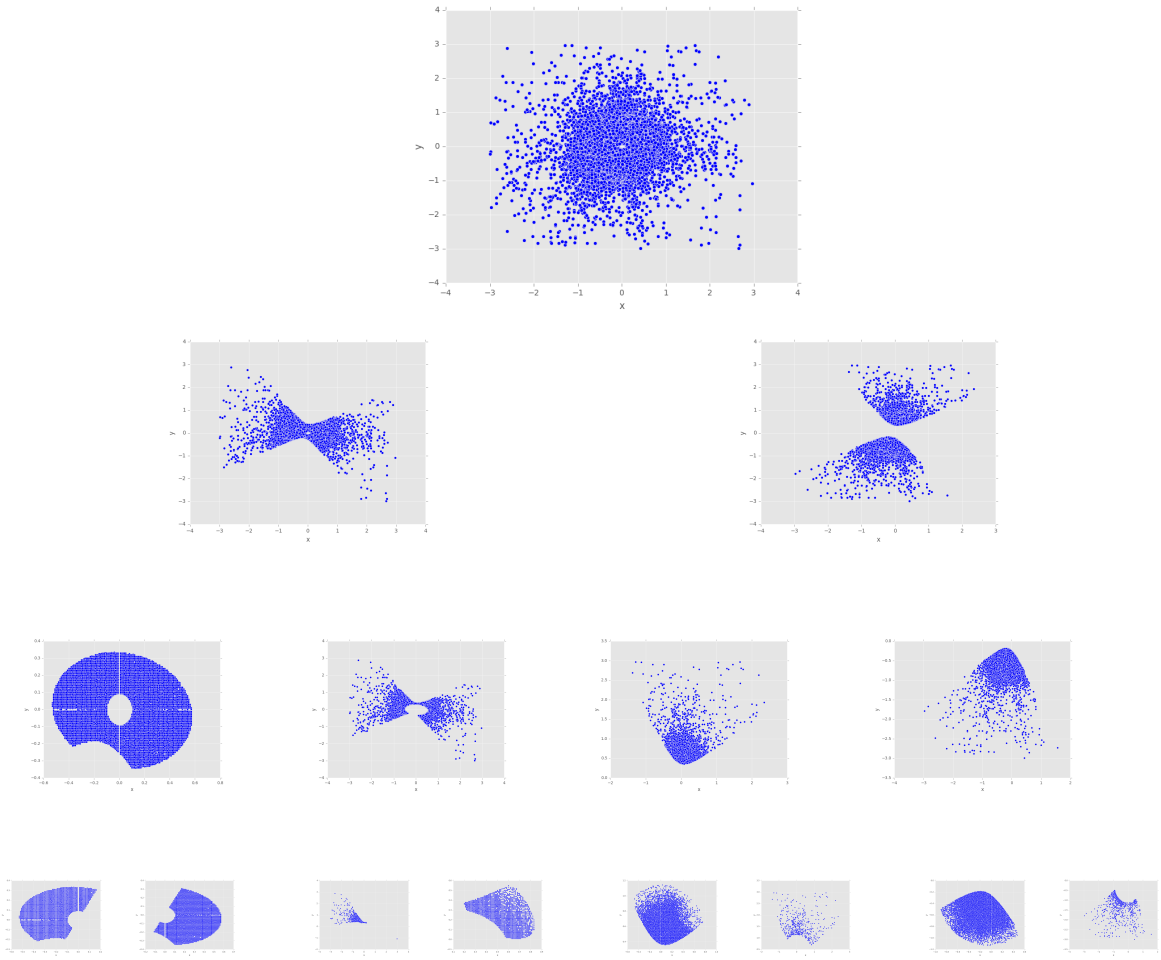


Figure 7.3: **Behavioural modules.** Movement modules across the first three layers of the hierarchical decomposition of the velocities of all sheep in the flock. It can be seen that the movement behaviour has been disentangled into modules corresponding to different behaviours of the sheep which, by extension, correspond to movement in different directions. These modules were further decomposed across three more hierarchical layers and the end module modelled with a mixture density network while each module was connected to its parent module using a recurrent neural network trained with the labels of the cluster at each level and the associated state values.

7.2.3 Training & Model details

As discussed earlier, a hierarchical decomposition of the movement behaviour was used in which I fit a two component (two cluster) Gaussian mixture model with full covariance type at each level of the hierarchy[6]. Overall, eight levels were used including six levels for the section of the architecture that outputs velocities, but the choice of the levels of the hierarchy is an open question and should be largely dependent on the size of the data and the number of Gaussian components that can be fitted to the data at any level of hierarchy. The decision to fit two components at several levels of hierarchy was borne out of the need to use a tree prior and, by extension, minimize the effect of misclassification as much as possible when deciding which of the movement modules to activate as the sensory inputs/stimuli move down the hierarchy. Making a choice between two things is relatively easy compared to choosing between many things. However, after conducting a number of initial experiments, I realized one of the modules at the third level of hierarchy downwards consisted of movement patterns where both right and left movement patterns were not disentangled properly but encoded together. Therefore, I used a four component Gaussian mixture model at the third level and continued with two afterwards. At the last layer of the hierarchy, I fit the most parsimonious number of Gaussian components by minimizing the Akaike information criterion (AIC). A discrete recurrent network was placed at each level of the hierarchy[7] except the final one to make a binary decision given state information. That is, during training and by extension during testing, these recurrent neural networks do the clustering work rather than the clustering algorithm so that when a sensory input is received, it can be processed, classified and propagated down the hierarchy.

Next, for the final layer of the hierarchy (the section that outputs velocities), I extracted for each sheep in the dataset the corresponding time where the action was taken and aggregated all the actions together. I further extracted the states at these points for each sheep where the states are as described above. I trained the different modules of behaviour by mapping the states to actions using the mixture density network characterised by a feed forward deep network with three layers, relu activation at each layer, drop-out = 0.2 at the final layer, learning rate = 1e-3 and an Adam optimizer [230]. These networks were trained until convergence.

All other networks were trained independently using a $90:10$ train to test ratio split using a recurrent neural network with a cross entropy loss and the weights corresponding to the epoch with the highest test

[6]Movement behaviour was considered from a collective perspective here where the velocities of all sheep in the flock were aggregated together.

[7]This network was trained using the cluster labels at each level and the associated states

accuracy selected. All parameters and architectures were selected via a few trial experiments but were majorly consistent across all networks with similar tasks. Further details about parameters and architectures can be found in Table 7.1. For the network making the decision to move or not, I trained and tested this network using data from five sheep drawn randomly from the flock. The parameters for the best performing (using accuracy as metric) sheep was then calibrated on the test data using only the three nearest neighbour state information.

Segment	**Description**
Collective behaviour recognition	LSTM, Optimization: Adam [230] (learning rate = 0.001), Batch size = 10, epochs = 100, activations = tanh, softmax, depth = 1 , time-step = 60secs.
Decision to move or not	LSTM, Optimization : Adam [230] (learning rate = 0.001), Batch size = 100, epochs = 100, activations = tanh, softmax, depth = 1, time-step = 5 secs.
Discrete hierarchical network	LSTM, Optimization: Adam [230] (learning rate = 0.001), Batch size = 10, Iterations = 200, activations = tanh, softmax, depth = 1, time-step = 5 secs.
Mixture Density Network	Feed forward network, Optimization: Adam [230] (learning rate = 0.001), Batch size = 1, Iterations = trained until convergence, activation = relu, layers = 3, 256 neurons in each layer.
%hline	

Table 7.1: **Details of configurations and parameters of architecture.**

To prevent the inefficiency of neural networks for practical and complex tasks [252], I refrained from training the whole network end to end. Instead, each neural network was trained apiece and stacked in a hierarchical structure akin to training one network at a time in a complex network/settings while freezing the rest. In other words, I used a series of if/else statements [253] for connecting adjacent networks together.

7.3 Materials & Data Processing

7.3.1 Animals & Data

Data for this experiment were collected for 11 days between 21/03/2016 and 08/04/2016. GPS data were collected using bespoke GPS loggers, details of which can be found in [200]. These were attached to the back of the sheep as shown in (see Figure 7.4). The GPS data were sampled at a rate of 1 sample/s with an estimated average positional accuracy of $\pm 10cm$[8] to ensure all forms of interesting movement patterns by

[8]Standard deviation of the actual position.

the sheep were captured. The datasets were all roughly 24 hours in length. Each dataset contains a phase in which loggers were attached to each sheep in a holding pen followed by a phase in which the sheep were herded into the field and then a phase in which the sheep were left to roam across the field and a final phase in which the sheep were herded back to the holding pen to have the logging device removed and re-charged. All computations were largely based on the data obtained when the sheep were left to wander in the field on their own as this represents the period in which their natural behaviour was most likely to be observed. The flock used for this study consists of 36 sheep that were moved around four locations across 11 days with an average period of 3 days in each location. As at the time of carrying out these experiments, no notable disorders have been found in these sheep[9].

7.3.2 Data Processing

Raw GPS data were combined with data from a local fixed base-station and processed using the open-source RTKLib and Geographic Library to obtain Cartesian co-ordinates with respect to a local projection. All missing data were interpolated between the last and next seen co-ordinates using a Kalman smoother whose parameters were tuned using Expectation Maximization algorithm. I removed the location data for one sheep in the training set as it has lots of missing data. For the experiments, I extracted the portion of the data where the animals were on the field until at least one of the GPS loggers attached to the sheep failed due to battery exhaustion to reduce the uncertainty inherent in estimating missing data to the extent possible. Overall, four days worth of data were extracted for training and testing, details of which can be found below.

7.3.2.1 Training Dataset

The training dataset was collected on 01/04/2016 and after processing contains about 16 hours of behavioural data. Approximately, the dimension of the field where the training data was collected is 142 x 136m.

7.3.2.2 Test Datasets

The proposed approach was evaluated on three datasets of the same flock collected on either different days or different location/field with respect to the training data or both.

Test Dataset 1: The dataset was collected on 21/03/2016 from an experiment conducted in a different location with respect to the training dataset with the size of the field used being approximately 73 x 66m. After preprocessing, the dataset consists of 16 hours of behavioural data.

[9]The flock is a mixture of normal and transgenic sheep.

Test Dataset 2: This dataset was collected from the same location as the training dataset on 04/04/2016 and after preprocessing consists of about 18 hours of behavioural data.

Test Dataset 3: The dataset was collected on 05/04/2016 from an experiment conducted in a different location with respect to the training dataset as well as test datasets 1 and 2. Approximately, the dimension of the field used for the experiment was 228 x 137m and after preprocessing, the dataset consists of about 19 hours of behavioural data.

Figure 7.4: **Sheep with GPS loggers attached to their back.**

I assumed all sheep in the training and test datasets were homogeneous to create more training data and diversity of samples for the sensory inputs with respect to the neural networks most especially for those that determine what module to select for velocity[10]. Therefore, the relationship between the inputs and outputs have been considered from a collective behaviour perspective. Each sheep had 57610 data points. After removing all instances where the collective behaviour was largely stationary and each sheep was not moving, there were about 491454 data points left for all sheep in the flock. More than two thirds of these data-points were approximated as zero after the first clustering as they were essentially noise. All actions for the sheep in the training dataset were extracted and aggregated into a two column vector representing the changes in x and y co-ordinates between contiguous points for further downstream tasks such as clustering.

[10]Data augmentation.

7.4 Experiments & Procedures

7.4.1 Baselines

I compared the proposed approach with three competitive baselines, details of which can be found below. Other baselines such as the Variational Recurrent Neural Network [254] and its conditional variant [255] were considered initially, but results show they are not competitive for complex tasks such as considered here.

7.4.1.1 Null Movement

I used a baseline with no movement to see how the model compares in this regard considering sheep are animals that are less mobile.

7.4.1.2 Recurrent Neural Network (LSTM)

The recurrent neural network with a many to one architecture and a mean square error training objectives was used to predict the velocities in a multitask fashion (see Figure 7.5). I used a 2-layer LSTM network each with 60 memory cells and a many to one architecture designed to multitask where the output of the LSTM layers was fed to two other layers with dropout (0.2) to predict the x and y speeds.

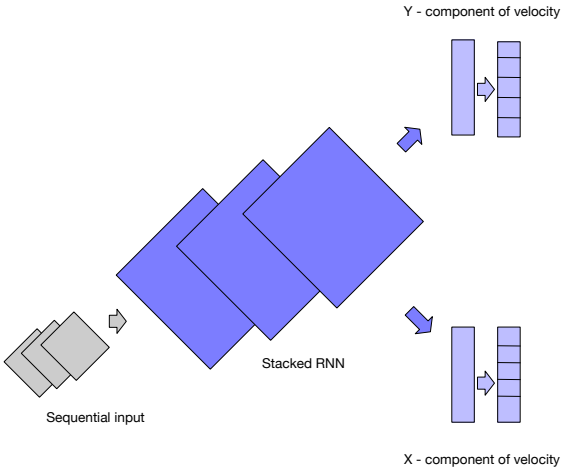


Figure 7.5: **RNN-LSTM.** Configured to multitask the prediction of the components of the velocities.

7.4.1.3 Boids

Even though some of the objectives of this work is to try and generate new hypotheses about the neural and cognitive basis of behaviour, I still compared it with a flocking model like boids. Boids [256] is an

artificial life program proposed to simulate the flocking behaviour of birds and is useful for studying emergent properties. It has three rules described below.

- Alignment: Move towards the average heading of local flock mates V_{ali}.
- Cohesion: Move towards the average position (centroid) of local flock mates V_{coh}.
- Separation: Steer clear to avoid overcrowding local flock mates. Here I used a radius threshold of $\leq$1m to decide when to steer clear V_{sep}.

While there are different versions of the boids model, I used the classical model because of its simplicity and ease of comparison[11]. The overall equation for an agent to update its velocity is thus defined as:

$$V_t = \beta_1 V_{ali} + \beta_2 V_{coh} + \beta_3 V_{sep} \tag{7.8}$$

The weights in the equation were learnt by optimization with respect to the target variables.

7.4.2 Evaluation

I used both a quantitative and qualitative evaluation approach to investigate the performance of the model proposed with respect to the baselines.

7.4.2.1 Quantitative evaluation (metrics)

To evaluate all methods quantitatively, I used three metrics described below.

- **Metric 1: N - Step average displacement error.** [257, 258] Measures the difference between the positions predicted by the models at each time step relative to the ground truth over N steps. Lower is better here. N steps of $1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096$[12] were considered where one step is equivalent to a second.

$$\frac{1}{M}\frac{1}{T}\sum_{m=1}^{M}\sum_{t=1}^{T}||y_{model_{mt}} - y_{ground-truth_{mt}}||_2^2 \tag{7.9}$$

The equation for this metric is described above, where M is the number of sheep in the flock and T the

[11] Some versions of boids tried, produced inconsistent behaviour across different nearest neighbours over different prediction horizons.

[12] These numbers were chosen to investigate performance over long horizons given the limit of the compute resources used.

temporal interval in consideration and is reset at $N, 2N, 3N.....$ steps[13] in that order and y_{model} as well as $y_{ground-truth}$ the x and y coordinates of the baseline models and the ground truth respectively.

- **Metric 2: N - Step final displacement error.** [258] Measures the difference between the final position after N steps and the corresponding ground truth. As with the previous metric, the lower the displacement error, the better.
- **Metric 3: N - Step average displacement divergence in distribution of velocity.** Here, I used the Jensen Shannon (JSD) divergence to find the distance between the output of the models and the ground truth using the distance covered as input. However, the probability density function, was estimated across the entire gamut of the data[14] so as to have access to a lot of data and by extension obtain a better probability density estimate. It is the case that, the lower the JSD, the better the model.

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{7.10}$$

Where P and Q are the respective probability density functions[15] of the two variables of interest and M is given by:

$$M = \frac{1}{2}(P + Q) \tag{7.11}$$

And the Kullback Leibler divergence $D(P||Q)$ is given by,

$$D(P||Q) = \int_{-\infty}^{+\infty} P(x)\ \log\ \frac{P(x)}{Q(x)}dx. \tag{7.12}$$

To test for significance where necessary, I used the **Mann Whitney U test**[16] with significance threshold of $\mathbf{p \leq 0.05}$.

[13] Reset because of compute resources constraint.
[14] By accumulating the distance covered over each N steps.
[15] The probability density functions were estimated using a top-hat kernel density estimator.
[16] A non-parametric approach used because the distribution of the errors is not Gaussian.

7.4.2.2 Qualitative evaluation

In addition, I also used a qualitative evaluation by inspecting and visualizing the trajectories obtained from the proposed approach in addition to some of the ablations.

7.4.3 Experiments

As usual, I designed several experiments to answer questions relating to the performance of the proposed approach with respect to the baselines as well as shed some light towards understanding some mechanisms underlying sheep movement.

7.4.3.1 Quantitative, Q1

How does varying the number of nearest neighbours state information influence movement activities? To answer this question, I computed the mutual information between nearest neighbour state variables in increasing order and velocity of the sheep of interest with respect to all animals in the flock and averaged the results.

7.4.3.2 Quantitative, Q2

How do the depth of the network/model as well as priors and assumptions used impact the performance of the model? To investigate this, I carried out several ablation studies, details of which can be found below.

7.4.3.3 Quantitative, Q3

How does the performance of proposed model compare to that of the baselines? To investigate this, I compared the best performing model from the ablation experiments in Q2 above with several baselines described above.

7.4.3.4 Qualitative, Q4

Are all hierarchies created equal? To investigate this, I used a qualitative evaluation to compare the performance of proposed approach and a naive implementation of this approach with an exact tree prior.

7.4.4 Ablation Studies

To investigate the performance of several priors and assumptions used for the proposed approach, I carried out several ablation experiments investigating three key features.

- Model 1: The layer predicting collective behaviour was removed and the network making the decision to move or not trained using the initial data used for this network and the data corresponding to that used for collective behaviour recognition.

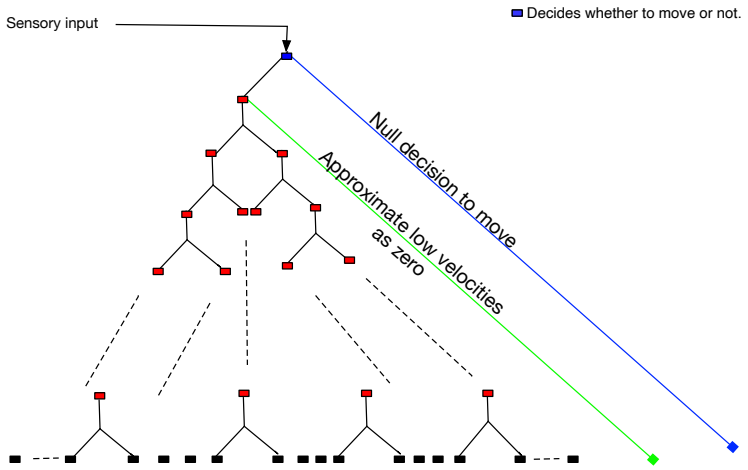


Figure 7.6: **Ablation study 1.** Architecture used for ablation study 1.

- Model 2: The depth of the network corresponding to the segment that output velocities was reduced from 6 to 3 to assess the effect of depth on the performance of the architecture.

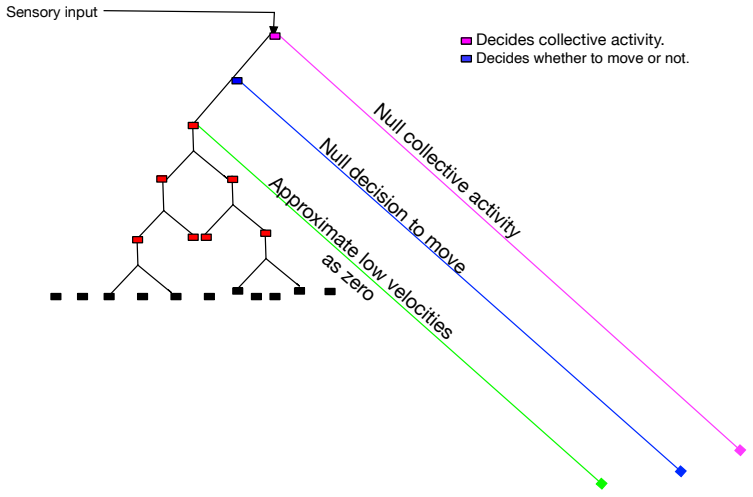


Figure 7.7: **Ablation study 2.** Architecture used for ablation study 2.

- Model 3: All the top control layers predicting both collective behaviour and the decision to move or not were removed to assess the impact of these layers on the performance of the architecture.

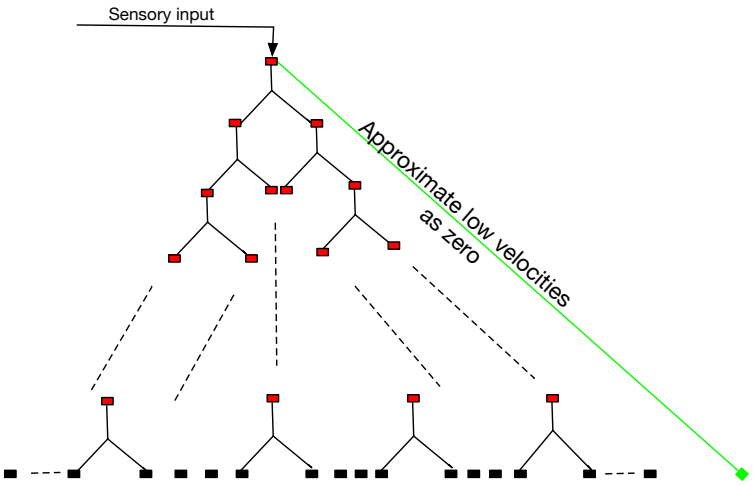


Figure 7.8: **Ablation study 3.** Architecture used for ablation study 3.

7.5 Results & Discussion

I discuss first the quantitative results in the order listed in the experiments section above followed by the qualitative results.

7.5.1 Quantitative Results

7.5.1.1 A1

The results of the experiments discussed in Q1 above were plotted with results in Figure 7.9 showing that the mutual information peaks at two nearest neighbours and monotonically decreases afterwards. This suggests that movement decisions are more local than global and that the movement of a minimum of two nearby sheep may be required before a sheep can decide to move.

7.5.1.2 A2 - Ablation Results

- First, the collective behaviour prior can be seen to be an important layer in the organization of behaviour. The plots in Figures 7.11 and 7.12 show that the inclusion of this prior gives results for metrics 1 & 2 that are significantly better than for the case in which it was omitted. The third metric has not been

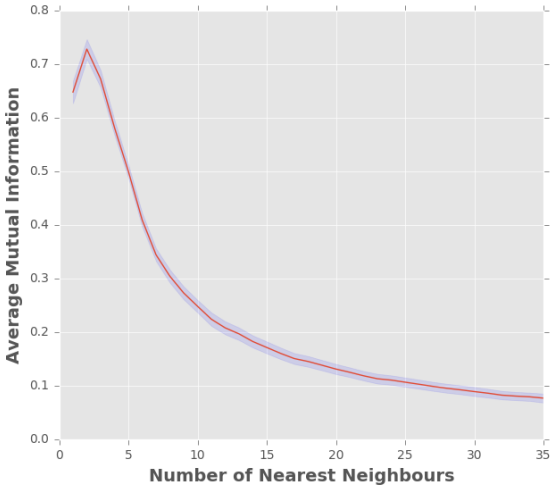


Figure 7.9: **Average mutual information between the features used and the velocities across all sheep in the flock.** It can be seen that the average mutual information rises and peaks at two to three nearest neighbours then falls monotonically afterwards.

considered in this section as the first two metrics are enough to evaluate performance. Qualitative evaluation in Figure 7.10 also show the trajectory with the collective behaviour prior is slightly better than without it suggesting this prior gives a better control of movement than without it.

- Second, the depth of the hierarchy used is important. The deeper the hierarchy, the more accurate the output of the model according to the evaluation metrics 1 and 2 (see Figures 7.13 and 7.14). This makes sense as fine grained behaviour is better represented using a deeper hierarchy compared to a shallow one. Qualitative evaluation in Figure 7.10 also show the trajectory with the reduced depth of hierarchy produces a trajectory of lower quality compared to when the depth has not been reduced. This is because of lack of fine-grained velocity information which can easily get the artificial agent into states that have not being seen during training.

- Lastly, Figures 7.15 and 7.16 show that all the top control layers are not necessarily important every-time and can be redundant sometimes simply because the low velocities movement behaviour has already been approximated as zero. Therefore, this serves as a form of control when the top layers[17] are removed. Qualitative evaluation in Figure 7.10 also show the trajectory is better in quality using a simpler hierarchical framework.

To this end, I therefore used this variant of the proposed architecture while comparing

[17]Which may incur extra limitations with attendant downstream consequences on the output of the model.

with other baselines.

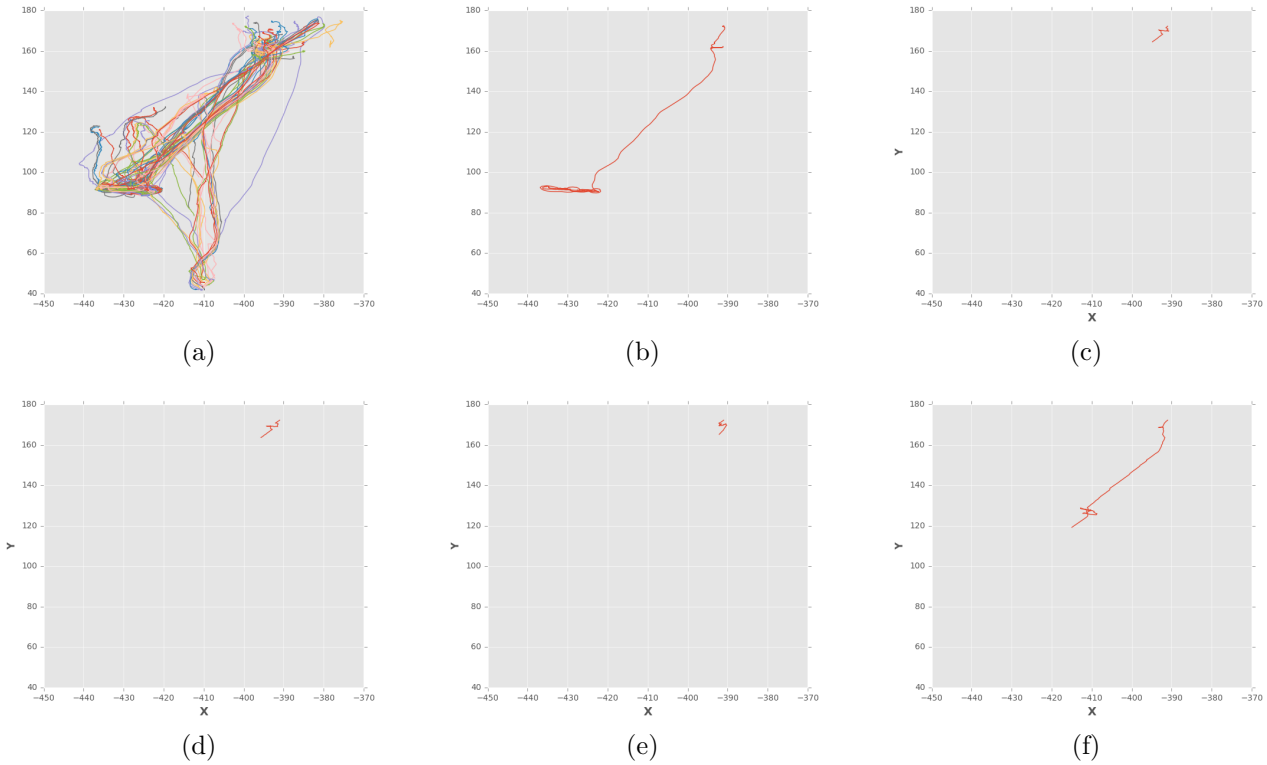


Figure 7.10: **Trajectories of ablations using three nearest neighbours state information over a certain period.** (a) Ground-truth trajectories of the flock. (b) Ground-truth trajectory of the sheep being modelled. (c) Output of the model with all the priors included. (d) Output of ablation - Model -1 (No collective behaviour recognition). (e) Output of ablation - Model -2 (Reduced depth of the hierarchy). (f) Output of ablation - Model -3 (Without top layer control).

7.5.1.3 A3 - Performance Comparison

- First, I compared the best performing model from A2 above with the null movement baseline. Results in Figures (7.17 and 7.18) show the proposed approach outperforms this baseline almost all of the time with performance decreasing as the number of nearest neighbours increases.
- Second, I compared the best performing model from A2 above with the recurrent neural network baseline. Results in Figures (7.19 and 7.20) show that the proposed approach perform better than a basic recurrent neural network almost all of the time. Similar to the no movement baseline above, the RNN outputs almost zero velocity most of the time. I hypothesize that this is due to its inability to handle more complex distributions as obtained in the data used here thus, the model's weights were adjusted such that they work for only instances where the animal doesn't move as this is the case most

of the time for sheep. In addition, the result also implies the priors/biases induced in the proposed model leveraging concepts of modularity and compositionality which can be weak in regular recurrent networks may be useful for modelling highly complex sheep movement behaviour.

- Third, I compared the best performing model from A2 above with boids. Results in Figures (7.21 and 7.22) show that the proposed approach outperform boids model at all nearest neighbours over short horizons but start to weaken over longer horizons. The boids model however appears like the most competitive among the baselines used with respect to the two metrics for evaluation.
- Finally, the results from the measure of divergence in the distribution between the output of the models and the ground truth show that most of the time, the proposed approach outperform all baselines across all nearest neighbours and prediction/simulation horizon over all experiments (Figure 7.23). This suggests that the movement of the simulated sheep using the proposed approach is more reflective of that of the real sheep relative to the distance covered at any point in time.

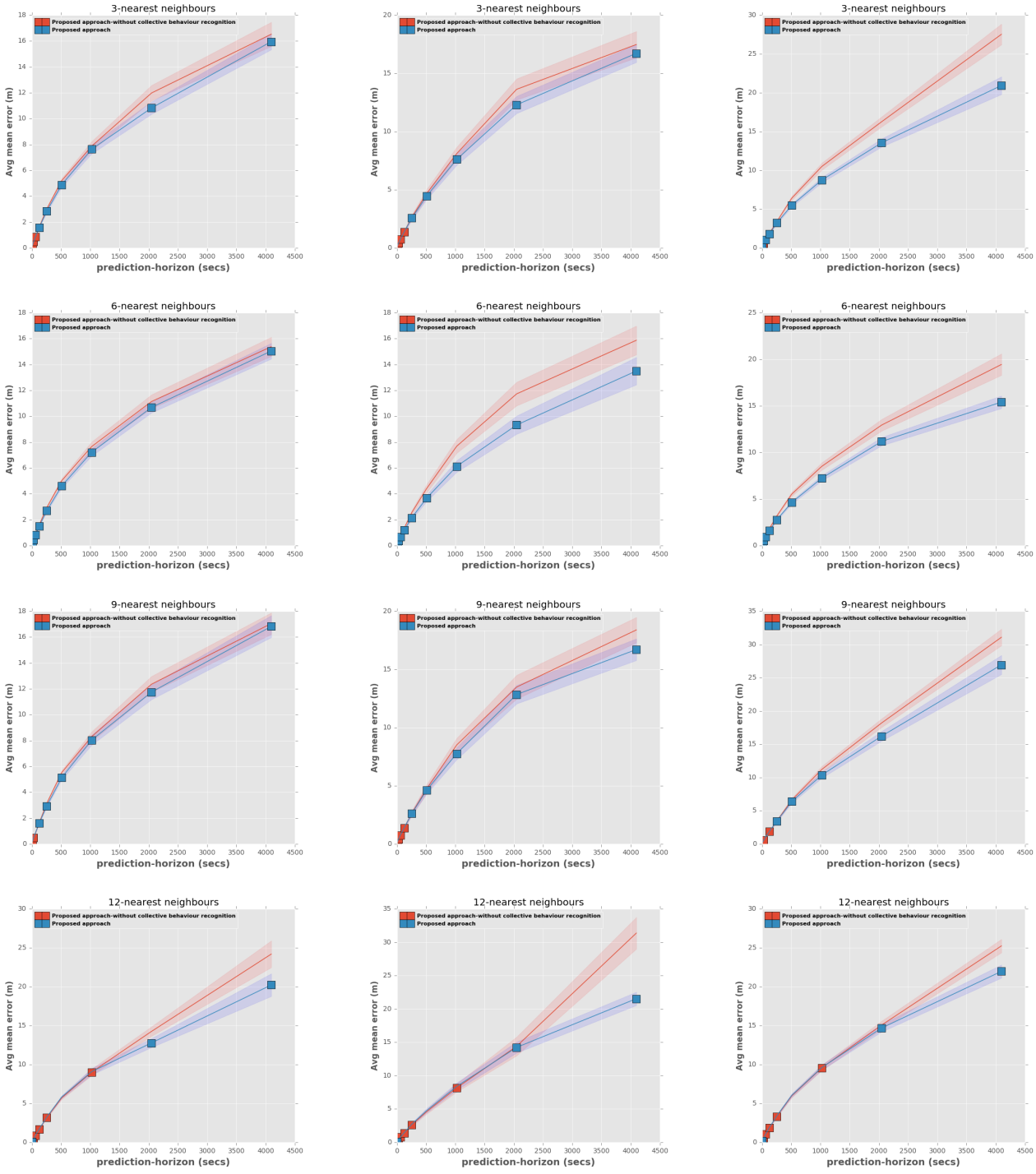


Figure 7.11: **Ablation - Model -1 (No collective behaviour recognition) - Metric-1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. It can be seen that the collective behaviour recognition is important for downstream tasks. The markers are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

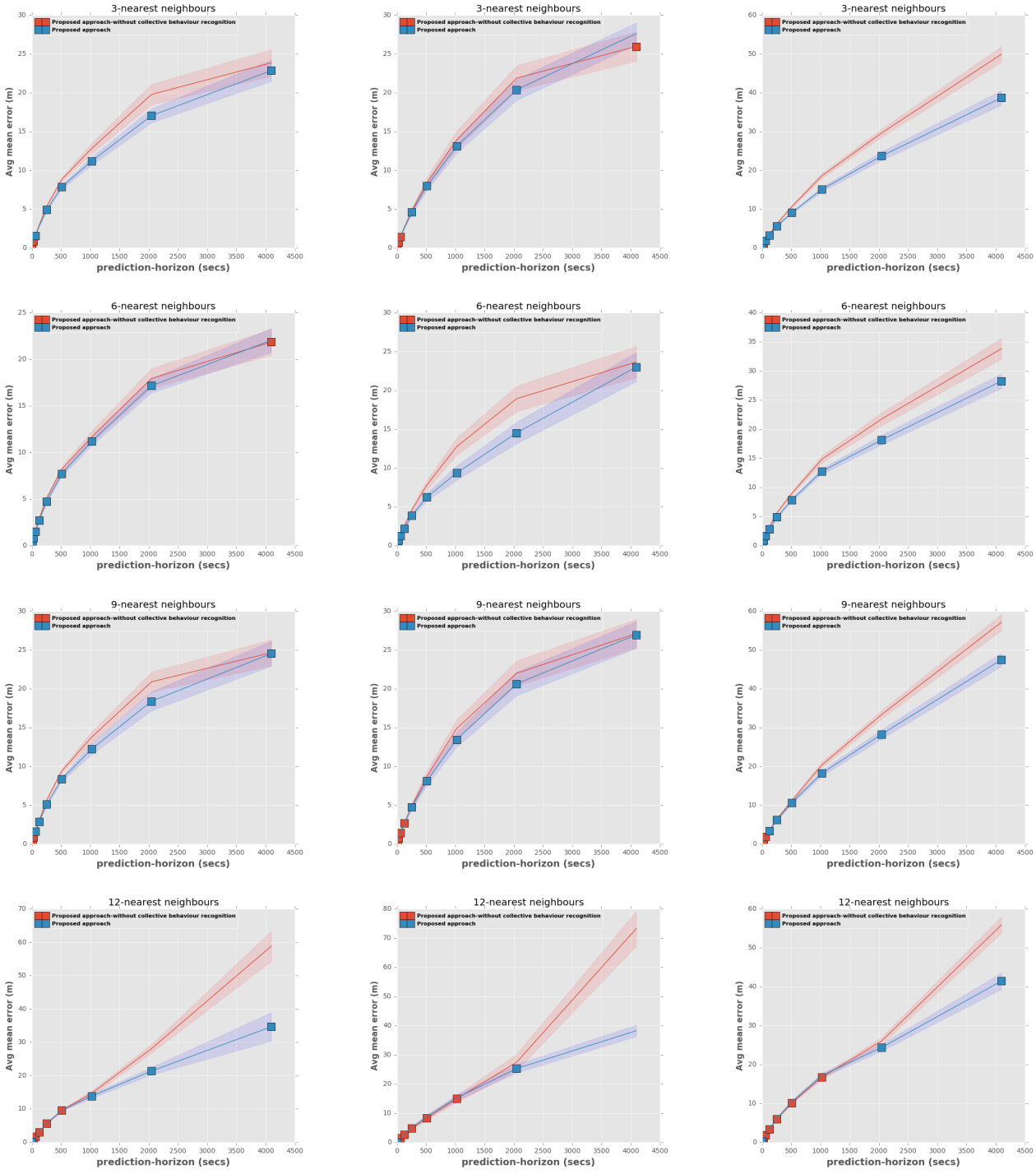


Figure 7.12: **Ablation - Model -1 (No collective behaviour recognition) - Metric-2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. It can be seen that the result degrades when collective behaviour is not used. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

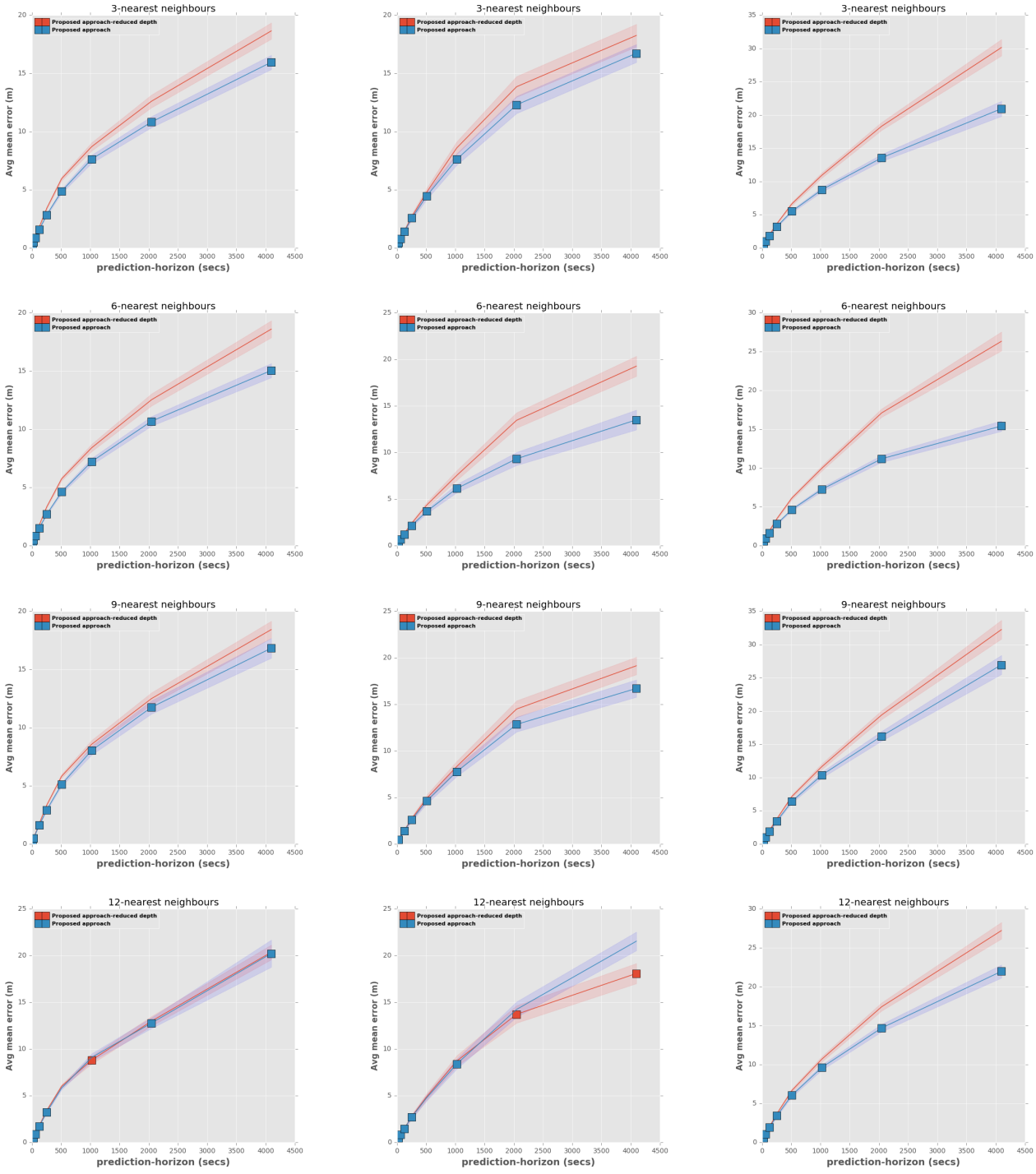


Figure 7.13: **Ablation - Model -2 (Reduced depth of the hierarchy) - Metric-1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. It can be seen that deep networks produce better results than a shallow one. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

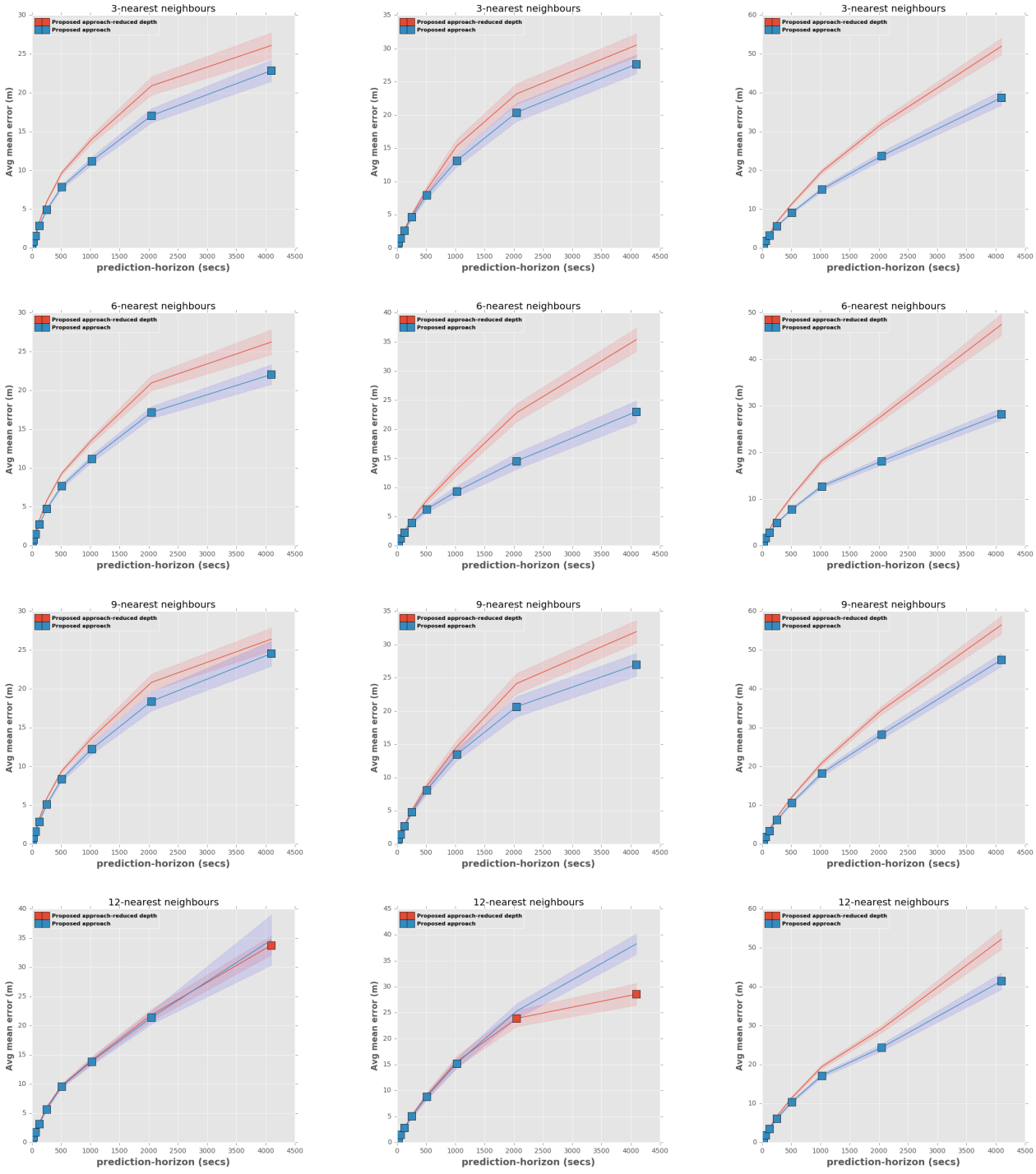


Figure 7.14: **Ablation - Model - 2 (Reduced depth of the hierarchy) - Metric-2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. It can be seen that the deeper the network the better the results. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

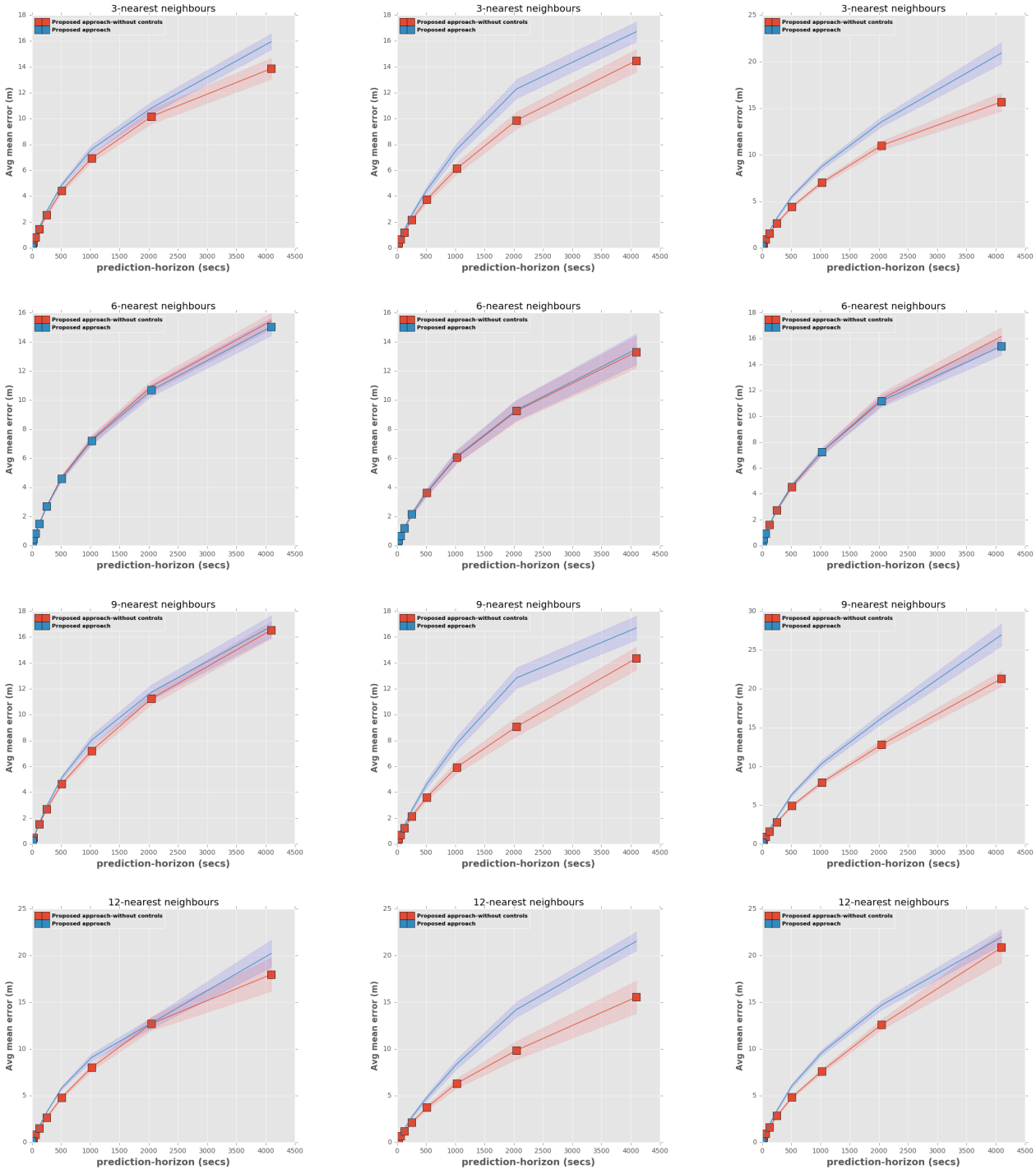


Figure 7.15: **Ablation - Model - 3 (Without top layer control) - Metric-1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. The proposed approach can be seen to perform better than this model sometimes and some other times the ablation model performs better. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

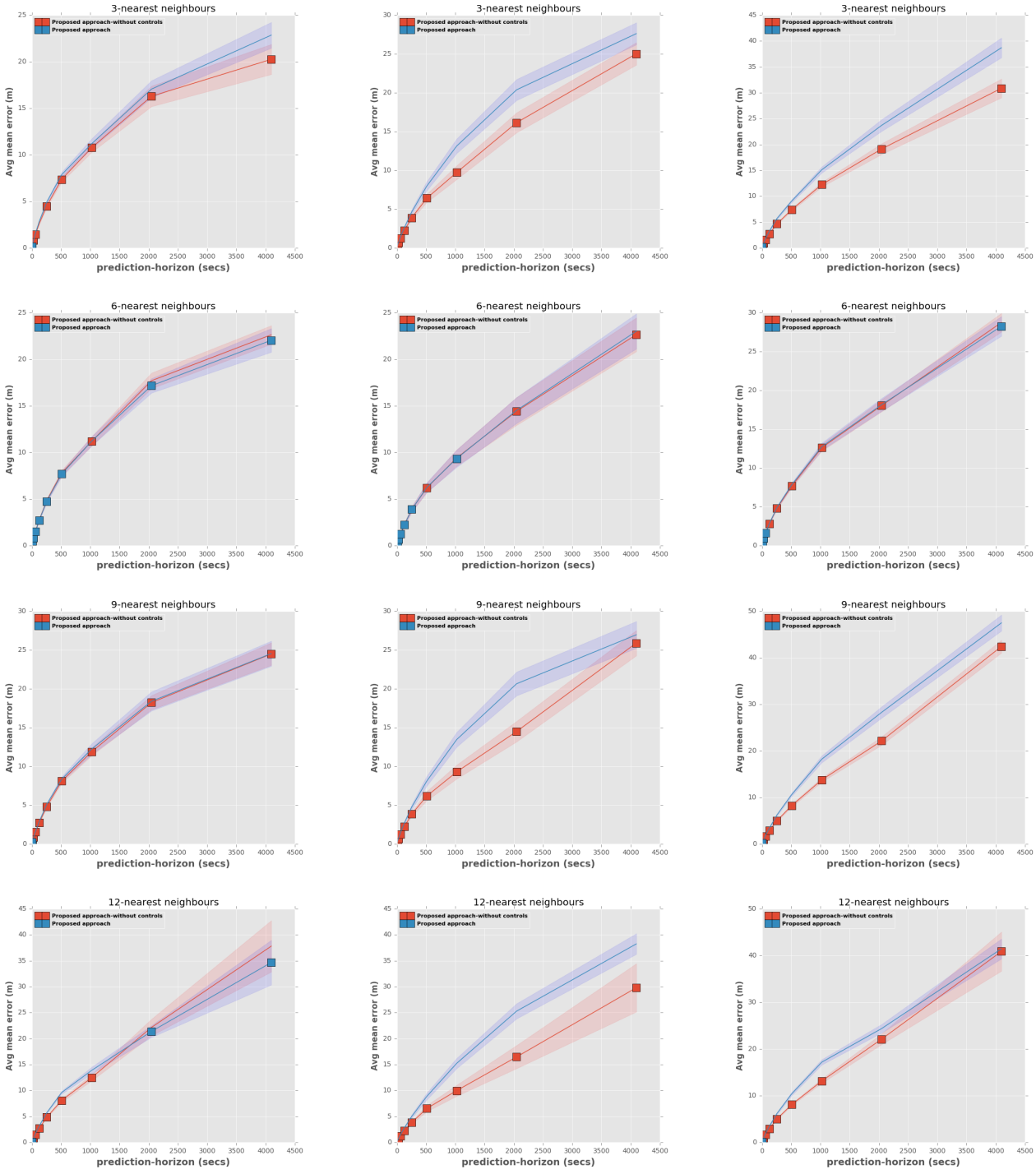


Figure 7.16: **Ablation - Model - 3 (Without top layer control) - Metric-2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. The proposed approach can be seen to perform better than this model sometimes and some other times the ablation model performs better. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

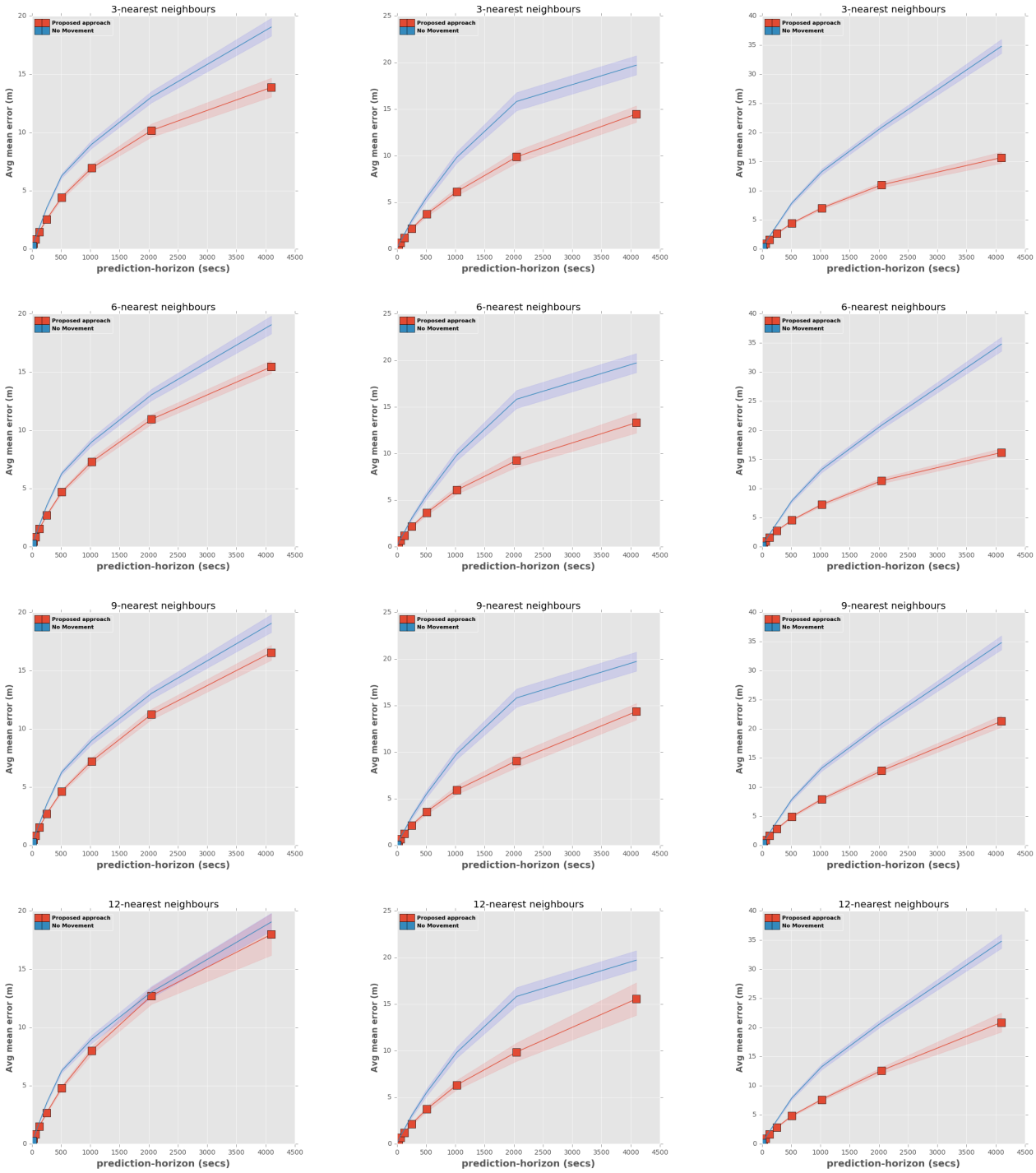


Figure 7.17: **Performance comparison with respect to baseline 1 - no movement using metric-1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. Proposed approach can be seen to outperform this baseline across different nearest neighbours and prediction-horizons. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

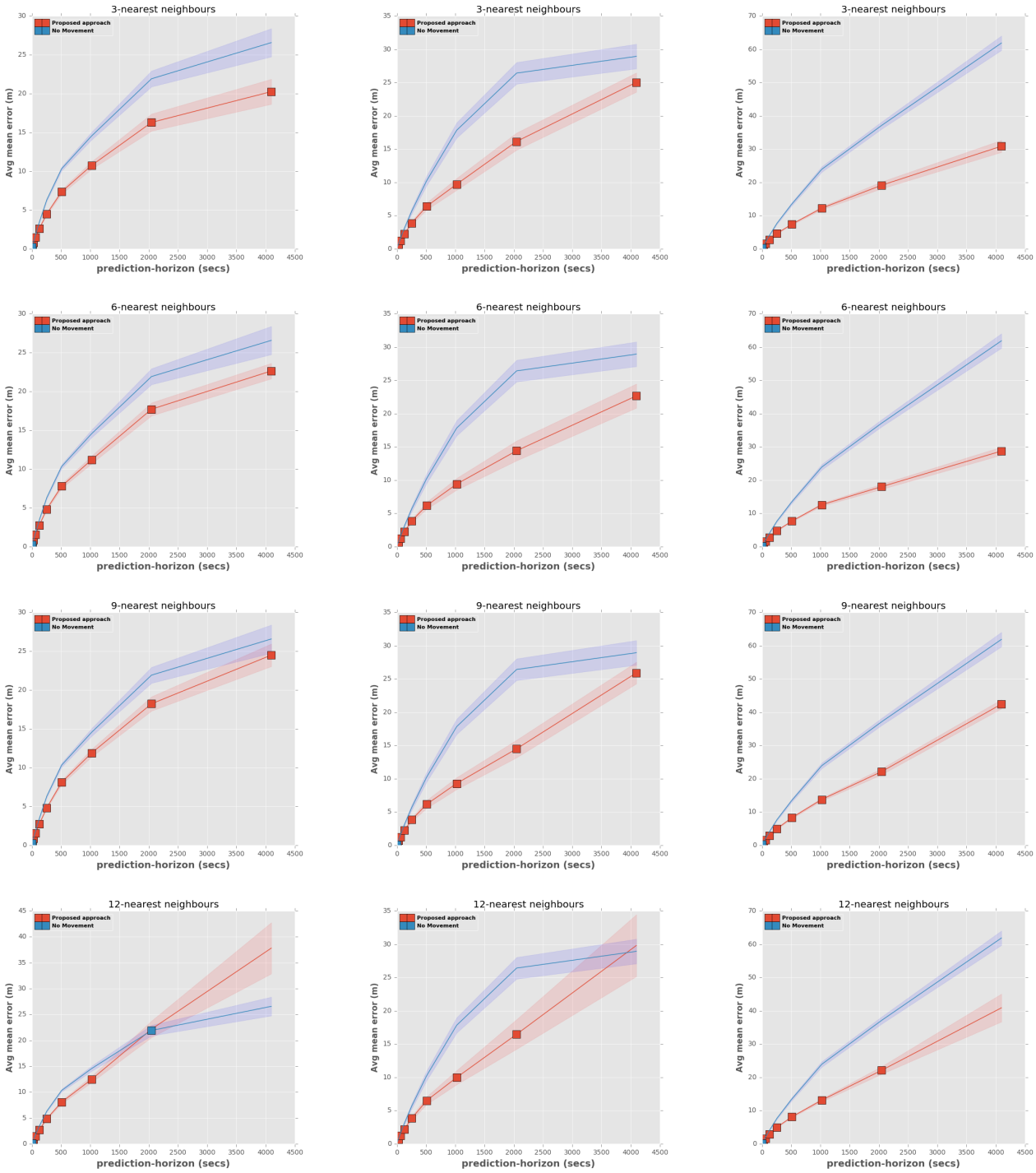


Figure 7.18: **Performance comparison with respect to baseline 1- no movement using metric-2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. As in metric-1, the proposed approach can be seen to outperform this baseline most of the time across all experiments. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

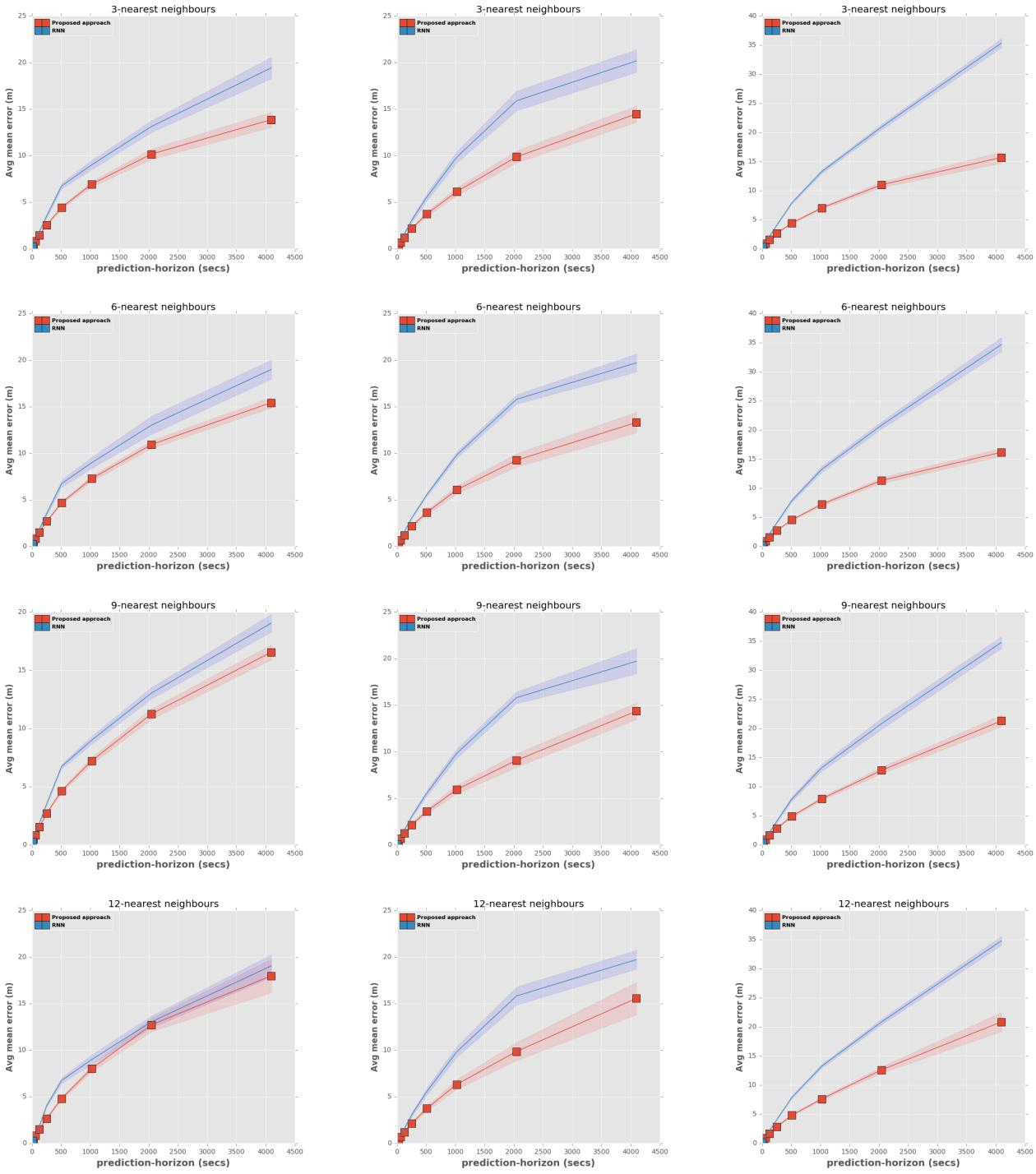


Figure 7.19: **Performance comparison with respect to baseline 2 - RNN using metric-1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. Proposed approach can be seen to outperform the RNN baseline almost across all experiments with respect to all nearest neighbours and prediction horizons. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

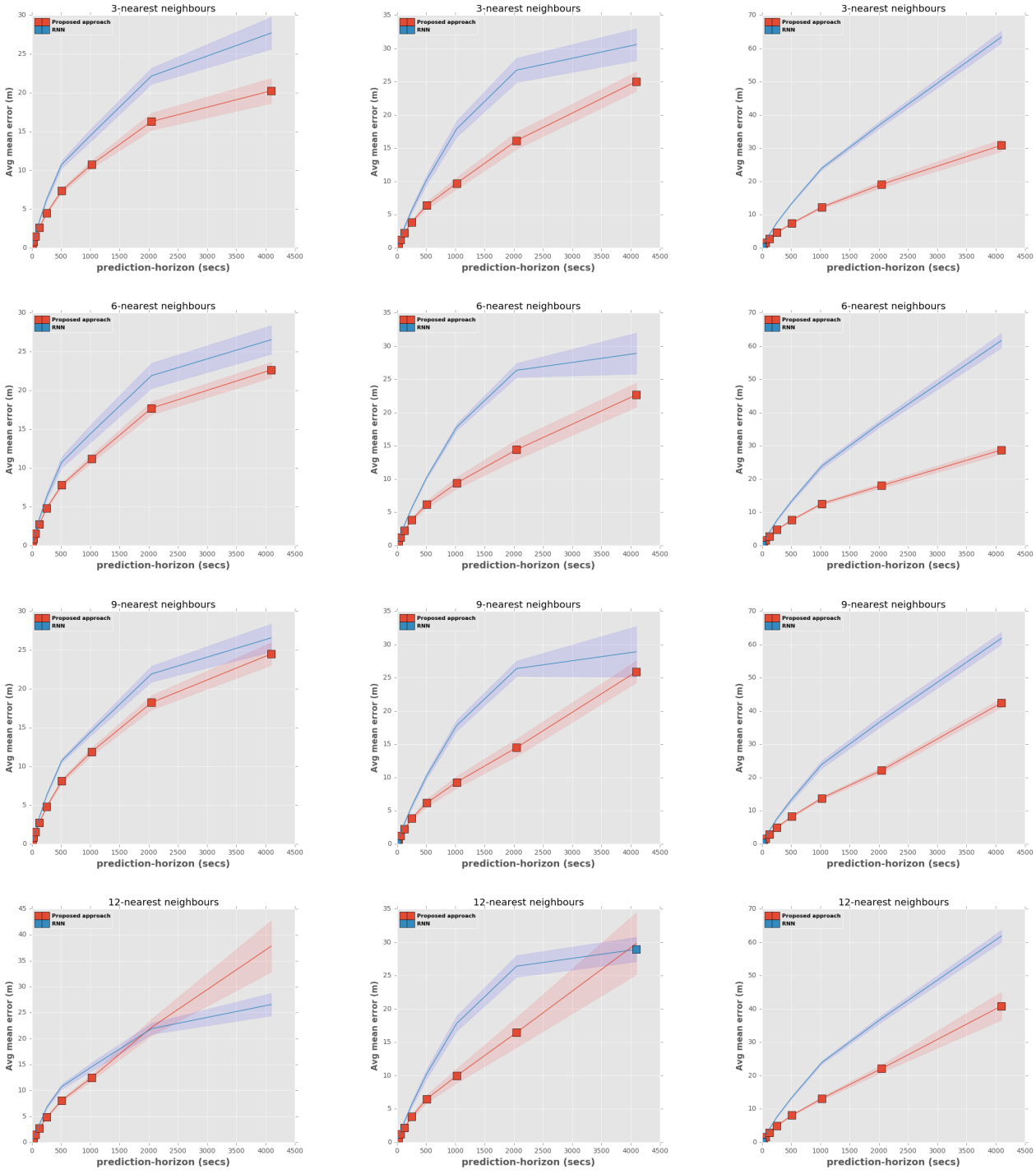


Figure 7.20: **Performance comparison with respect to baseline 2 - RNN using metric-2.** Left: Test dataset 1, Centre: Test dataset, 2 Right: Test dataset 3. Proposed approach can be seen to outperform the RNN most of the time across all experiments. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

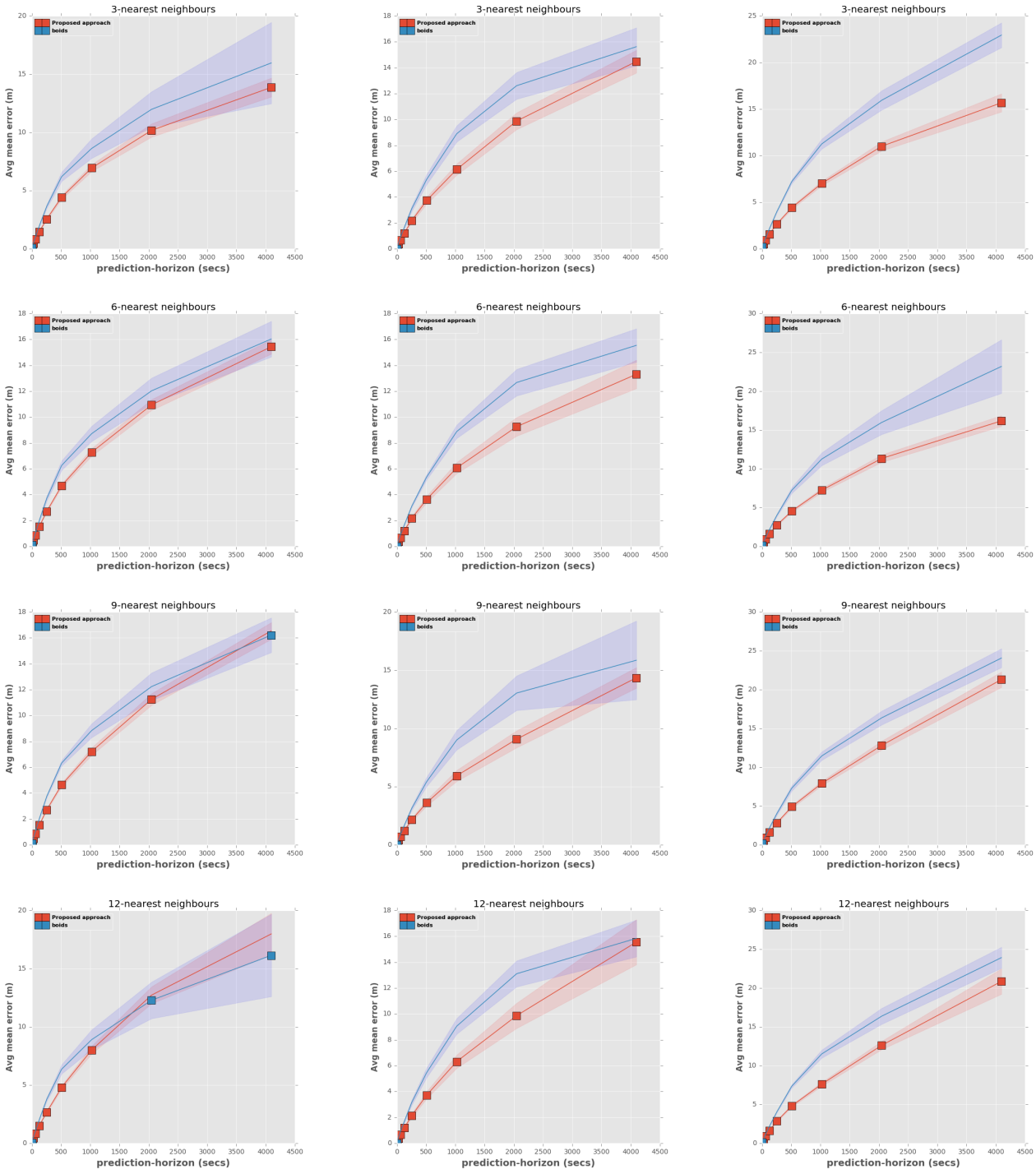


Figure 7.21: **Performance comparison with respect to baseline 3 - boids using metric-1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. Proposed approach can be seen to outperform boids most times when the number of nearest neighbours is small with performance decreasing as the number of nearest neighbours increases. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

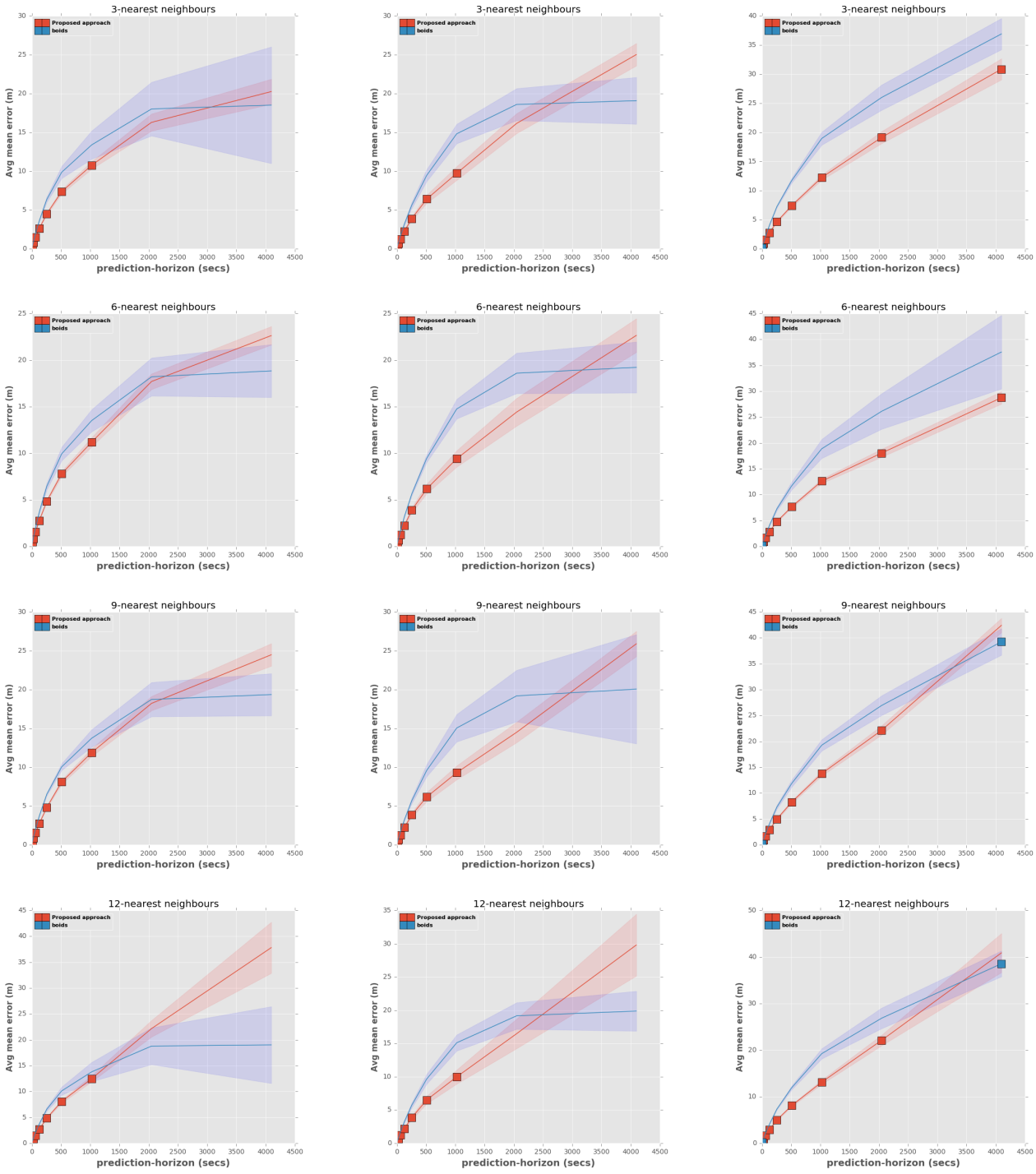


Figure 7.22: **Performance comparison with respect to baseline 3 - boids using metric-2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. Proposed approach can be seen to outperform boids most of the time across all experiments with performance decreasing with increase in simulation/prediction horizon as well as nearest neighbours. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test. See Appendix C for more results.

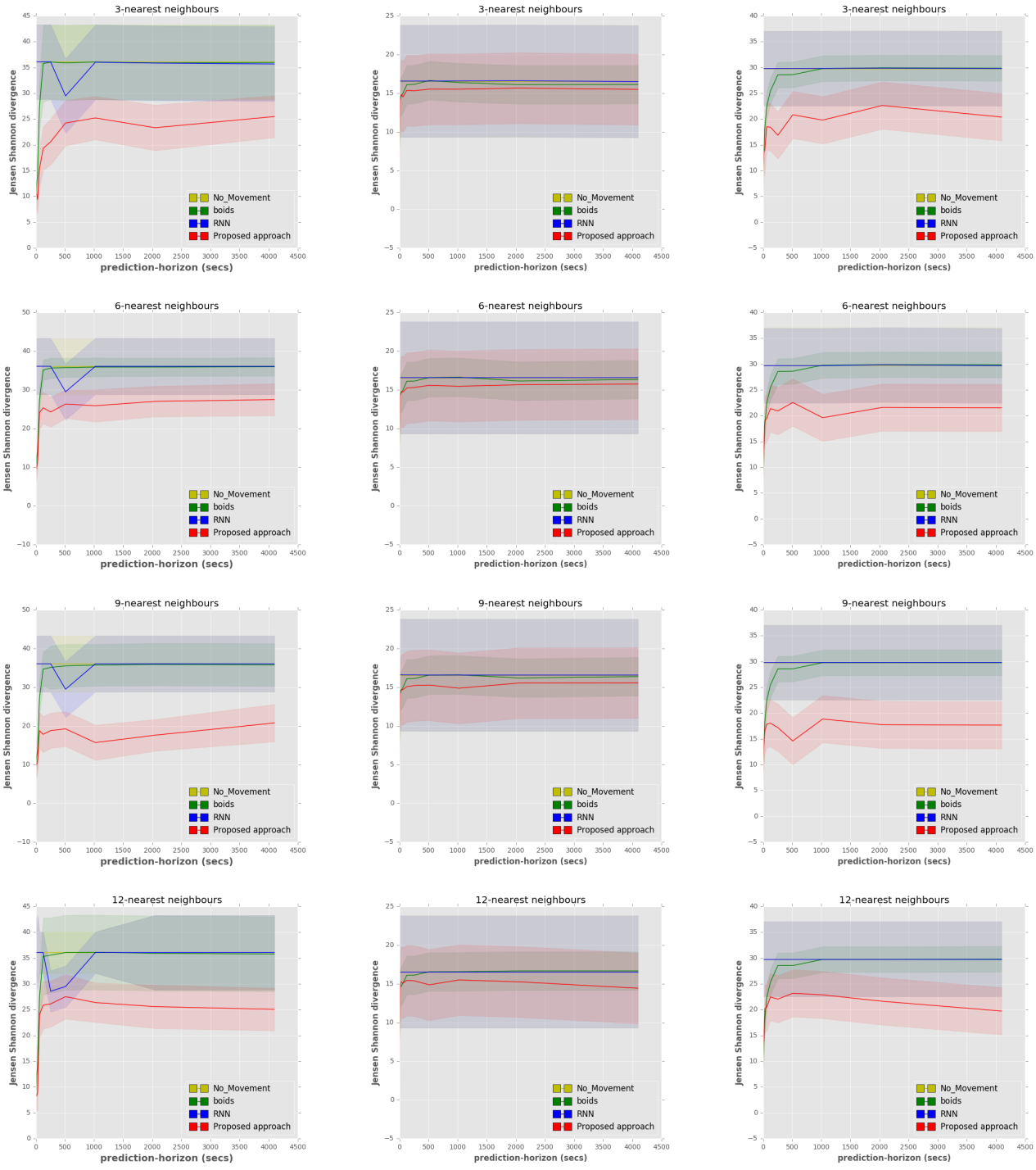


Figure 7.23: **Performance comparison with respect to all baselines using metric - 3.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. It can be seen that the proposed approach outperforms all baselines almost all the time across all the experiments. See Appendix C for more results.

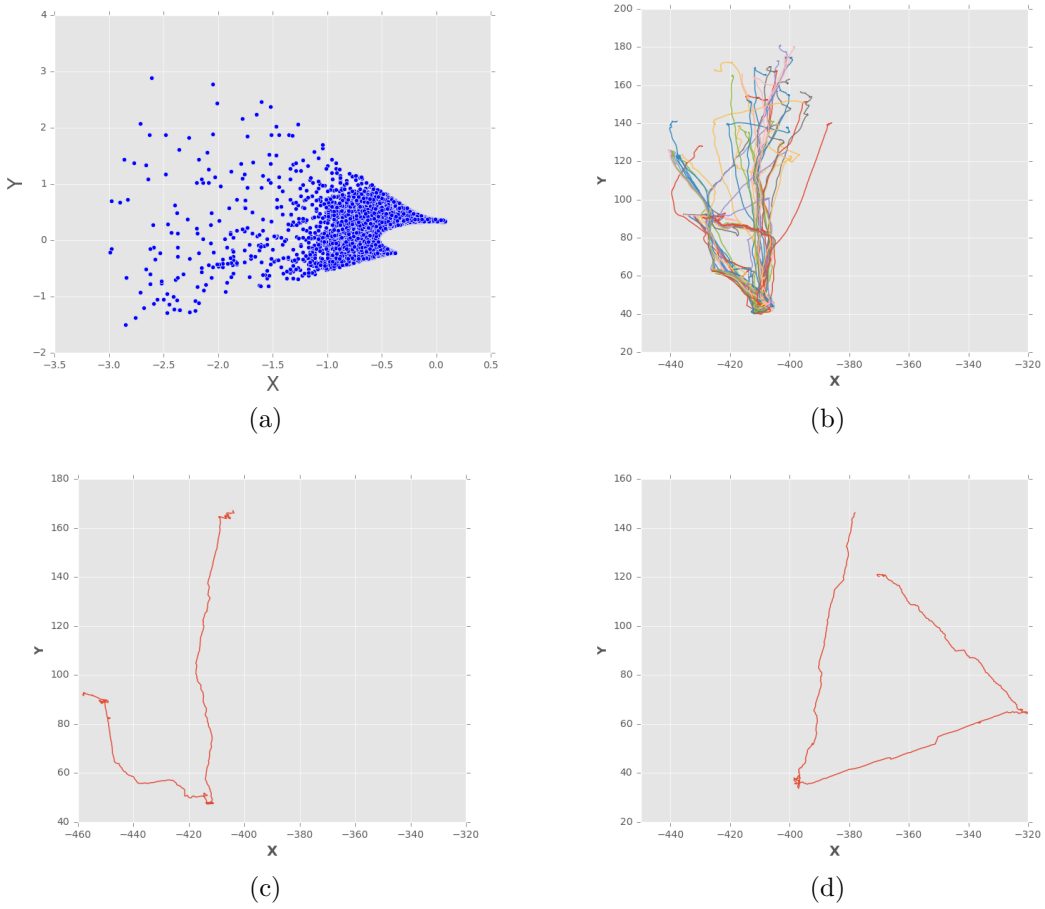


Figure 7.24: **Qualitative evaluation.** (a) Anomalous movement module (representation) encoding both left and right movement decisions. (b) Movement trajectories of the whole flock (c) Movement trajectory of the simulated agent using the proposed approach with correction for the anomalous representation. (d) Movement trajectory of the simulated agent using a naive implementation of the proposed approach with exact tree prior. It can be seen that the artificial agent can sometimes fail to distinguish between left and right movement decisions as they are encoded together when a naive implementation of the proposed approach was used, moving right or left sometimes when the module above was selected.

7.5.2 Qualitative Results

7.5.2.1 A4

Results in Figure 7.24 show that a naive implementation of the proposed architecture can yield movement patterns that are not sheep like as modules corresponding to unrelated movement patterns can be disentangled together leading to unusual behaviour sometimes. This is a problem, however, that might be solved with more data/experience. But, in this case, this could be an issue that concerns the optimality of the hierarchy

for representation.

I also evaluated the model by inspecting the trajectories of the simulations amidst the real flock. Details can be found in [18], [19], [20], [21], [22] for a single agent simulation and, multiagent here[23]. The roll-outs[24] were generated by picking random points in the test dataset while learning the movement patterns from there onwards where the red agent is the simulated version of the blue agent (when present). Observations reveal the artificial sheep has learned to herd and also behave somewhat like a sheep by following and responding to the real sheep around it. The trajectory for the non-naive implementation (Figure 7.24C) can also be seen to have several curves implying that the artificial sheep has learned to change direction of movement where necessary and not just follow a straight line.

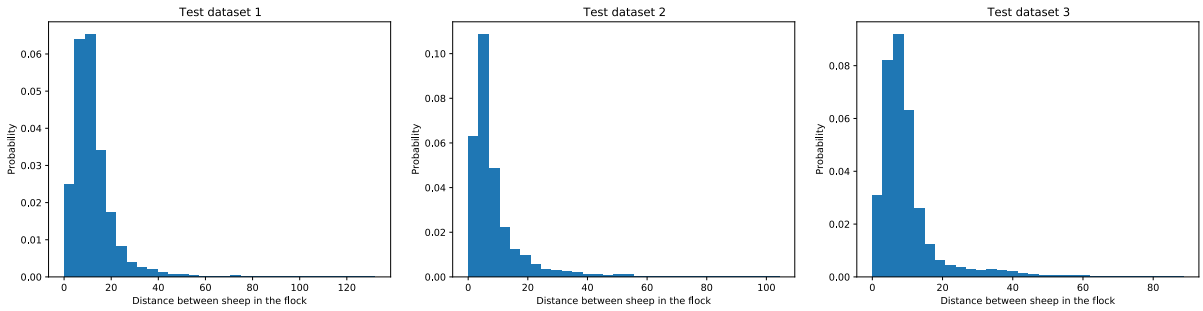


Figure 7.25: **Ground truth distribution of the pairwise distance in metre between all the sheep in the flock for the test datasets.**

7.6 Conclusion & Limitations

In this work, I have proposed a model leveraging modularity, compositionality and relational reasoning principles to build a novel hierarchical deep learning model towards modelling sheep movement behaviour in a flock. This was achieved by decomposing behaviour into modules hierarchically and using the interconnection between these modules to predict the movement of sheep in a flock over long horizons with results performing better than the baselines most of the time.

[18] https://www.youtube.com/watch?v=QUu3ejxDQt4
[19] https://www.youtube.com/watch?v=kYQ7HoXDb1M&t=99s
[20] https://www.youtube.com/watch?v=psfVQAfIr-U&t=3s
[21] https://www.youtube.com/watch?v=7F1gYQuqkTg
[22] https://www.youtube.com/watch?v=hbSr1gAhS_Q
[23] https://www.youtube.com/watch?v=0okpGFi61_M&t=116s
[24] The behaviour simulated exceed three hours and good behaviour of more than seven hours have been seen. The problem with simulating very good behaviour over very long horizons has to do with lack of the right quality of state information such as vision and the approximation of the low velocities as zero.

From the experiments and ablations carried out, it can be observed that:

- Hierarchical priors may be important and most importantly, hierarchies of sequences and features may be important for animal movement behaviour modelling.
- The deeper this hierarchy is, the better the behaviour representation and the output of the proposed model.
- As a consequence of the above, multiple behaviour representations may exist at multiple scales of the hierarchy both temporally and spatially.
- Discrete and continuous representations may be important for modelling complex sheep movement behaviour.
- And most importantly, the meaningful organization and segmentation of behaviour may be important for modelling sheep movement behaviour.
- Qualitative evaluation also reveals herd like behaviour in the artificial sheep.

Quantitatively, the magnitude of displacement error for the proposed architecture is, first a function of the size of the environment in which the experiments were carried out. That is, the larger the size the bigger the error as the animals have the freedom to explore their environment. Second, a function of the duration of these experiments with respect to the test data implying that, the longer the duration the higher the magnitude of displacement error. Third, is within the limit of what is expected to be the distance between sheep in the flock on the average (Figure 7.25) suggesting the artificial sheep is able to herd even when not behaving exactly like the ground truth sheep its being compared to. Lastly, a function of the number of nearest neighbours, with the magnitude of displacement error increasing as the number of nearest neighbours increases.

Qualitatively, while a particular number of nearest neighbours state information can be optimal for some behaviour, it may not be so for others suggesting rich state input such as data collected from ego-centric vision can yield better results. As a consequence of this, the nearest neighbours state information used here means, the magnitude of the final displacement error is larger compared to the average displacement error because there is a lag in the start of several movement decisions by the artificial sheep compared to the ground-truth. In addition, the trajectory for the non-naive implementation (Figure 7.24C) can also be seen to have several curves implying that the artificial sheep has learned to behave intelligently and not just follow a shortcut straight line.

This work however is not without its limitations. The quality of data used here in terms of the sensory inputs and the associated motor output here can benefit from additional quality. Rich representations of the sensory inputs such as what the animal can see using ego centric vision can alleviate this limitation most especially for the deep learning models. Due to the number of networks used, the same network parameters have been used for all networks performing similar tasks. Careful investigation of different parameters for each network might yield superior results. Finally, while the approach used here does incur additional computational costs compared to other models for example like boids, it can be advantageous where visual sensory data corresponding to what the animal can see have been collected [84][25]. In addition, it can also be used outside of animal behaviour modelling for example in human behaviour modelling where richer representations of the environment/sensory inputs are obtainable and not limited to just information about conspecifics (this will be useful in several areas of behaviour data mining such as sport analytics, assisted living, self driving vehicles among others where recent advances in neural networks can be leveraged towards obtaining more utility from the relevant data of interest).

In this chapter, I have demonstrated the novel combination of both supervised and unsupervised methods towards modelling and predicting sheep movement behaviour in a flock. In the next chapter, I will conclude this thesis and highlight avenues for future work.

[25]In this scenario, boids or other simple mathematical models cannot be used.

Chapter 8

Conclusion & Future work

8.1 Conclusion

In this thesis, several computational models of animal movement behaviour have been proposed towards addressing a wide range of problems in animal movement phenomena ranging from migration, welfare, leadership and association, understanding neural mechanisms underlying behaviour among others. More specifically:

- In chapter three, I reviewed and investigated the application of various information theoretic metrics towards understanding and finding patterns in animal movement behaviour.
- In chapter four, several statistical and machine learning models of abnormal behaviour in sheep with Batten disease have been proposed and compared. Extensive experiments show these models can be used to diagnose the occurrence of phenotypes associated with abnormal movements in this sheep.
- In chapter five, a deep recurrent neural network leveraging auxiliary task have been proposed towards forecasting several migration patterns in Turkey vulture. Results on a variety of experiments show the performance of the proposed approach is superior to a variety of baselines.
- In chapter six, I proposed a model to forecast stopover decisions of migratory birds where a pseudo loss objective was proposed with promising results.
- In chapter seven, model of sheep movement behaviour in a flock incorporating reasoning modules have been proposed towards predicting and understanding the structure of movement behaviour in a flock

of sheep. Results show that the proposed approach outperformed several baselines most times on a variety of metrics.

The models proposed here have implications in several areas such as cognitive neuroscience, global health, conservation and sustainable development just to mention a few. And, it is hoped that the findings and the models proposed will be useful to researchers and policy makers in these fields.

There are several limitations associated with the contributions in this thesis and have been discussed at the end of each relevant chapter. In general however, a bulk of the limitations centre around the quantity of animals used for some of the experiments for example in chapters 5&6, and I aim to follow up on that in the future. Also, the performance gains of the models proposed in these chapters though marginal sometimes, especially over longer horizons for the model proposed in chapter 6 for instance, I aim to improve on that in the future.

8.2 Future Work

There are several promising future directions with respect to this body of work and related ones in the following areas:

- **Identifying the onset of symptoms in sheep with Huntington disease:** Huntington disease similar to Batten disease is a genetic neurodegenerative disorder. While preliminary work suggests there aren't too much differences if any in the movement patterns of sheep with this disease relative to their control group, recent study carried out [198] suggests there are significant differences in the metabolic activities between the two groups of sheep. We therefore plan to mine the long-term accelerometer and movement data of these sheep that we possess to see if there are significant changes in behaviour and if any to know when they actually started starting with some of the methods proposed in this thesis [259].
- **Memory and integration of information:** While human and perhaps animal brain can integrate information about an event or a course of action over a period of time, it is not clear how to build machine learning systems with this capability. One major key to achieving this will however be memory networks and investigating architectures motivated by the behaviour of living systems will be key towards unlocking the mechanism behind this capability.
- **Understanding mechanisms for continual learning in animals:** Similar to the above, to build the next generation of animal inspired robotics and artificial intelligence, understanding the mechanisms

used by these animals to update knowledge will be important. To this end, several directions with respect to continual learning in animals are worth investigating. These include but not limited to: What priors if necessary are important? How are new information updated? How do these animals decide what to remember and what to forget? What kind of neural architectures are optimal for these kind of tasks?

- **Evaluating the performance of anomaly detection algorithms without synthetic data:** While I have evaluated the performance of the anomaly detection methods using carefully generated synthetic data, this may not be ideal. In the future, developing metrics useful for evaluating the performance of these models will be important given just the training data.
- **Compositional behaviour reasoning:** While a supervised hierarchical approach has been used to select relevant behavioural module given a sensory input in this work, the question is how can these modules be used to reason compositionally about highly novel situations?
- **Optimal hierarchy:** While I have shown in this thesis, hierarchies might be important components of information storage and retrieval in the brain of animals, learning how to automatically design the optimal hierarchy is a promising direction towards ensuring the right information is quickly and efficiently recovered at little computational cost.
- **Learning with few labelled data:** Collecting real world behaviour data is without any doubt expensive. And even more expensive is the labelling of this data for downstream application tasks in understanding animal behaviour. While self-supervised learning has been suggested as a solution to learning with few labels, it is not obviously clear how this can generalise to non-stationary distributions with multiple modalities. Investigating the use of self supervised learning in deep sequential generative model is therefore a direction I aim to consider in future work.
- **Augmenting information in forecasting tasks:** The mutual information between inputs and targets appears to decrease albeit at different rates for different tasks as the forecasting horizon increases [260, 261]. This automatically leads to a decrease in performance for models as this information decreases. One way to reduce this effect could be to augment the input data in such a way that the information loss decreases at a slower rate as the forecast horizon decreases whilst not compromising the amount of information necessary to predict the targets successfully and I aim to investigate this in. the future.

Bibliography

[1] Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8):649–663, 2005.

[2] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanislaw Jastrzebski, Thibault Fevry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists? performance in breast cancer screening. *IEEE transactions on medical imaging*, 2019.

[3] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, pages 1–5, 2020.

[4] Charles K Fisher, Aaron M Smith, and Jonathan R Walsh. Machine learning for comprehensive forecasting of alzheimer?s disease progression. *Scientific reports*, 9(1):1–14, 2019.

[5] Rory P Wilson. Estimation of location: global coverage using light intensity. *Wildlife telemetry: remote monitoring and tracking of animals*, 1992.

[6] Philina A English, Alexander M Mills, Michael D Cadman, Audrey E Heagy, Greg J Rand, David J Green, and Joseph J Nocera. Tracking the migration of a nocturnal aerial insectivore in the americas. *BMC Zoology*, 2(1):1–11, 2017.

[7] Anders P Tøttrup, Lykke Pedersen, Alejandro Onrubia, Raymond HG Klaassen, and Kasper Thorup. Migration of red-backed shrikes from the iberian peninsula: optimal or sub-optimal detour? *Journal of Avian Biology*, 48(1):149–154, 2017.

[8] WMG Vansteelant, J Kekkonen, and P Byholm. Wind conditions and geography shape the first outbound migration of juvenile honey buzzards and their distribution across sub-saharan africa. *Proceedings of the Royal Society B: Biological Sciences*, 284(1855):20170387, 2017.

[9] Jaewon Sa, Miso Ju, Seoungyup Han, Heegon Kim, Yongwha Chung, and Daihee Park. Detection of low-weight pigs by using a top-view camera. In *Proc. of The Fourth International Conference on Information Science and Cloud Computing*, volume 24, 2015.

[10] Y Chung, J Lee, S Oh, D Park, HH Chang, and S Kim. Automatic detection of cow?s oestrus in audio surveillance system. *Asian-Australasian journal of animal sciences*, 26(7):1030, 2013.

[11] Jonguk Lee, Long Jin, Daihee Park, and Yongwha Chung. Automatic recognition of aggressive behavior in pigs using a kinect depth sensor. *Sensors*, 16(5):631, 2016.

[12] Jonguk Lee, Byeongjoon Noh, Suin Jang, Daihee Park, Yongwha Chung, and Hong-Hee Chang. Stress detection and classification of laying hens by sound analysis. *Asian-Australasian journal of animal sciences*, 28(4):592, 2015.

[13] Natasha K Boyland, David T Mlynski, Richard James, Lauren JN Brent, and Darren P Croft. The social network structure of a dynamic group of dairy cows: From individual to group level patterns. *Applied Animal Behaviour Science*, 174:1–10, 2016.

[14] Henri Weimerskirch, David Pinaud, Frédéric Pawlowski, and Charles-André Bost. Does prey capture induce area-restricted search? a fine-scale study using gps in a marine predator, the wandering albatross. *The American Naturalist*, 170(5):734–743, 2007.

[15] Oliver Padget, Gaia Dell?Ariccia, Anna Gagliardo, Jacob González-Solís, and Tim Guilford. Anosmia impairs homing orientation but not foraging behaviour in free-ranging shearwaters. *Scientific reports*, 7 (1):1–12, 2017.

[16] Juliet S Lamb, Yvan G Satgé, and Patrick GR Jodice. Influence of density-dependent competition on foraging and migratory behavior of a subtropical colonial seabird. *Ecology and evolution*, 7(16): 6469–6481, 2017.

[17] David J Augustine and Justin D Derner. Assessing herbivore foraging behavior with gps collars in a semiarid grassland. *Sensors*, 13(3):3711–3723, 2013.

[18] Hall Sawyer, Matthew J Kauffman, and Ryan M Nielson. Influence of well pad activity on winter habitat selection patterns of mule deer. *The Journal of Wildlife Management*, 73(7):1052–1061, 2009.

[19] Jesse Whittington, Colleen Cassady St. Clair, and George Mercer. Spatial responses of wolves to roads and trails in mountain valleys. *Ecological Applications*, 15(2):543–553, 2005.

[20] Simon Thirgood, Anna Mosser, Sebastian Tham, Grant Hopcraft, Ephraim Mwangomo, Titus Mlengeya, Morris Kilewo, John Fryxell, ARE Sinclair, and Markus Borner. Can parks protect migratory ungulates? the case of the serengeti wildebeest. *Animal Conservation*, 7(2):113–120, 2004.

[21] Clément Calenge, Stéphane Dray, and Manuela Royer-Carenzi. The concept of animals' trajectories from a data analysis perspective. *Ecological informatics*, 4(1):34–41, 2009.

[22] Charles B Yackulic, Stephen Blake, Sharon Deem, Michael Kock, and María Uriarte. One size does not fit all: flexible models are required to understand animal movement across scales. *Journal of Animal Ecology*, 80(5):1088–1096, 2011.

[23] Kirsten E Ironside, David J Mattson, Tad Theimer, Brian Jansen, Brandon Holton, Terence Arundel, Michael Peters, Joseph O Sexton, and Thomas C Edwards. Quantifying animal movement for caching foragers: the path identification index (pii) and cougars, puma concolor. *Movement ecology*, 5(1):1–17, 2017.

[24] Eliezer Gurarie, Christen H Fleming, William F Fagan, Kristin L Laidre, Jesús Hernández-Pliego, and Otso Ovaskainen. Correlated velocity models as a fundamental unit of animal movement: synthesis and applications. *Movement ecology*, 5(1):1–18, 2017.

[25] Navinder J Singh, Andrew M Allen, and Göran Ericsson. Quantifying migration behaviour using net squared displacement approach: clarifications and caveats. *PLoS One*, 11(3):e0149594, 2016.

[26] Simon Benhamou. How to reliably estimate the tortuosity of an animal's path:: straightness, sinuosity, or fractal dimension? *Journal of theoretical biology*, 229(2):209–220, 2004.

[27] Hendrik Edelhoff, Johannes Signer, and Niko Balkenhol. Path segmentation for beginners: an overview of current methods for detecting changes in animal movement patterns. *Movement ecology*, 4(1):1–21, 2016.

[28] Cedric Gaucherel. Wavelet analysis to detect regime shifts in animal movement. *Computational Ecology and Software*, 1(2):69, 2011.

[29] Maitreyi Sur, Andrew K Skidmore, Klaus-Michael Exo, Tiejun Wang, Bruno J Ens, and AG Toxopeus. Change detection in animal movement using discrete wavelet analysis. *Ecological informatics*, 20:47–57, 2014.

[30] Ali Soleymani, Frank Pennekamp, Somayeh Dodge, and Robert Weibel. Characterizing change points

and continuous transitions in movement behaviours using wavelet decomposition. *Methods in Ecology and Evolution*, 8(9):1113–1123, 2017.

[31] Eliezer Gurarie, Russel D Andrews, and Kristin L Laidre. A novel method for identifying behavioural changes in animal movement data. *Ecology letters*, 12(5):395–408, 2009.

[32] Rahel Noser and Richard W Byrne. Change point analysis of travel routes reveals novel insights into foraging strategies and cognitive maps of wild baboons. *American Journal of Primatology*, 76(5): 399–409, 2014.

[33] Alastair Franke, Terry Caelli, and Robert J Hudson. Analysis of movements and behavior of caribou (rangifer tarandus) using hidden markov models. *Ecological Modelling*, 173(2-3):259–270, 2004.

[34] Joanna Mills Flemming, Ian D Jonsen, Ransom A Myers, and Christopher A Field. Hierarchical state-space estimation of leatherback turtle navigation ability. *PLoS One*, 5(12):e14245, 2010.

[35] James D Forester, Anthony R Ives, Monica G Turner, Dean P Anderson, Daniel Fortin, Hawthorne L Beyer, Douglas W Smith, and Mark S Boyce. State–space models link elk movement patterns to landscape characteristics in yellowstone national park. *Ecological Monographs*, 77(2):285–299, 2007.

[36] Cigdem Beyan and Robert B Fisher. Detecting abnormal fish trajectories using clustered and labeled data. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 1476–1480. IEEE, 2013.

[37] Jung-Hua Wang, Shih-Kai Lee, Yi-Chung Lai, Cheng-Chun Lin, Ting-Yuan Wang, Ying-Ren Lin, Te-Hua Hsu, Chang-Wen Huang, and Chung-Ping Chiang. Anomalous behaviors detection for underwater fish using ai techniques. *IEEE Access*, 2020.

[38] Chunlei Xia, Longwen Fu, Zuoyi Liu, Hui Liu, Lingxin Chen, and Yuedan Liu. Aquatic toxic analysis by monitoring fish behavior using computer vision: A recent progress. *Journal of toxicology*, 2018, 2018.

[39] Anna-Lena Giesert, Wolf-Tilo Balke, and Gerhard Jahns. Probabilistic analysis of coughs in pigs to diagnose respiratory infections. *Landbauforschung-vTI Agriculture and Forestry Research*, 61(3): 237–242, 2011.

[40] Natalia Risi, Késia Oliveira Silva, Paulo RF Zulato, Rodrigo C Guido, and Giselle Borges. Use of artificial intelligence to identify vocalizations emitted by sick and healthy piglets. In *Livestock*

Environment VIII, 31 August–4 September 2008, Iguassu Falls, Brazil, page 140. American Society of Agricultural and Biological Engineers, 2009.

[41] Xiaolin Zhuang, Minna Bi, Jilei Guo, Siyu Wu, and Tiemin Zhang. Development of an early warning algorithm to detect sick broilers. *Computers and Electronics in Agriculture*, 144:102–113, 2018.

[42] Sofia Broomé, Karina Bech Gleerup, Pia Haubro Andersen, and Hedvig Kjellstrom. Dynamics are important for the recognition of equine pain in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12667–12676, 2019.

[43] Niek Andresen, Manuel Wöllhaf, Katharina Hohlbaum, Lars Lewejohann, Olaf Hellwich, Christa Thöne-Reineke, and Vitaly Belik. Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. *Plos one*, 15(4):e0228059, 2020.

[44] Alexander H Tuttle, Mark J Molinaro, Jasmine F Jethwa, Susana G Sotocinal, Juan C Prieto, Martin A Styner, Jeffrey S Mogil, and Mark J Zylka. A deep neural network to assess spontaneous pain from mouse facial expressions. *Molecular pain*, 14:1744806918763658, 2018.

[45] Yiting Lu, Marwa Mahmoud, and Peter Robinson. Estimating sheep pain level using facial action unit detection. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 394–399. IEEE, 2017.

[46] Nicholas Perentos, Amadeu Q Martins, Thomas C Watson, Ullrich Bartsch, Nadia L Mitchell, David N Palmer, Matthew W Jones, and A Jennifer Morton. Translational neurophysiology in sheep: measuring sleep and neurological dysfunction in CLN5 batten disease affected sheep. *Brain*, 138(4):862–874, 2015.

[47] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM*, 2020.

[48] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[49] Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & knowledge engineering*, 60(1):208–221, 2007.

[50] Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650, 2003.

[51] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework

for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002.

[52] Ji Zhang and Hai Wang. Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. *Knowledge and information systems*, 10(3):333–355, 2006.

[53] Nong Ye and Qiang Chen. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International*, 17(2):105–112, 2001.

[54] Peter J Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79, 2011.

[55] Kehinde Owoeye, Mirco Musolesi, and Stephen Hailes. Characterizing animal movement patterns across different scales and habitats using information theory. *bioRxiv*, page 311241, 2018.

[56] Kehinde Owoeye, Mirco Musolesi, and Stephen Hailes. Characterizing animal movement patterns across different scales and habitats using information theory. *bioRxiv*, 2018.

[57] Andreas Arning, Rakesh Agrawal, and Prabhakar Raghavan. A linear method for deviation detection in large databases. In *KDD*, volume 1141, pages 972–981, 1996.

[58] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[59] Amrudin Agovic, Arindam Banerjee, Auroop R Ganguly, and Vladimir Protopopescu. Anomaly detection in transportation corridors using manifold embedding. *Knowledge Discovery from Sensor Data*, pages 81–105, 2008.

[60] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680, 2014.

[61] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[62] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs.

Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[63] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1705–1714, 2019.

[64] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.

[65] T Trancart, A Acou, E De Oliveira, and E Feunteun. Forecasting animal migration using sarimax: an efficient means of reducing silver eel mortality caused by turbines. *Endangered Species Research*, 21(2): 181–190, 2013.

[66] Benjamin M Van Doren and Kyle G Horton. A continental system for forecasting bird migration. *Science*, 361(6407):1115–1118, 2018.

[67] J Van Belle, J Shamoun-Baranes, E Van Loon, and W Bouten. An operational model predicting autumn bird migration intensities for flight safety. *Journal of applied ecology*, 44(4):864–874, 2007.

[68] Kyle G Horton, Frank A La Sorte, Daniel Sheldon, Tsung-Yu Lin, Kevin Winner, Garrett Bernstein, Subhransu Maji, Wesley M Hochachka, and Andrew Farnsworth. Phenology of nocturnal avian migration has shifted at the continental scale. *Nature Climate Change*, pages 1–6, 2019.

[69] W John Richardson. Timing and amount of bird migration in relation to weather: a review. *Oikos*, pages 224–272, 1978.

[70] Ulya Bayram, Ruichen Sun, and William Lee. Modeling stopover sites of migratory birds ? routes for conservation of population and prevention of disease. 2017.

[71] J Amy Belaire, Betty J Kreakie, Timothy Keitt, and Emily Minor. Predicting and mapping potential whooping crane stopover habitat to guide site selection for wind energy projects. *Conservation biology*, 28(2):541–550, 2014.

[72] Andreas Kaiser. Stopover strategies in birds: a review of methods for estimating stopover length. *Bird Study*, 46(sup1):S299–S308, 1999.

[73] Mathieu Fourment, Aaron E Darling, and Edward C Holmes. The impact of migratory flyways on the spread of avian influenza virus in north america. *BMC evolutionary biology*, 17(1):118, 2017.

[74] A Marm Kilpatrick, Aleksei A Chmura, David W Gibbons, Robert C Fleischer, Peter P Marra, and Peter Daszak. Predicting the global spread of h5n1 avian influenza. *Proceedings of the National Academy of Sciences*, 103(51):19368–19373, 2006.

[75] Josanne H Verhagen, HP van der Jeugd, Bart A Nolet, Roy Slaterus, SP Kharitonov, PP de Vries, Oanh Vuong, Frank Majoor, Thijs Kuiken, and RA Fouchier. Wild bird surveillance around outbreaks of highly pathogenic avian influenza a (h5n8) virus in the netherlands, 2014, within the context of global flyways. *Eurosurveillance*, 20(12):21069, 2015.

[76] George Valiakos, Konstantinos Papaspyropoulos, Alexios Giannakopoulos, Periklis Birtsas, Sotirios Tsiodras, Michael R Hutchings, Vassiliki Spyrou, Danai Pervanidou, Labrini V Athanasiou, Nikolaos Papadopoulos, et al. Use of wild bird surveillance, human case data and gis spatial analysis for predicting spatial distributions of west nile virus in greece. *PLoS One*, 9(5):e96935, 2014.

[77] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[78] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

[79] Shoichiro Yamaguchi, Honda Naoki, Muneki Ikeda, Yuki Tsukada, Shunji Nakano, Ikue Mori, and Shin Ishii. Identification of animal behavioral strategies by inverse reinforcement learning. *PLoS computational biology*, 14(5):e1006122, 2018.

[80] Krishnamurthy Dvijotham and Emanuel Todorov. Inverse optimal control with linearly-solvable mdps. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 335–342, 2010.

[81] Josh Merel, Diego Aldarondo, Jesse Marshall, Yuval Tassa, Greg Wayne, and Bence Ölveczky. Deep neuroethology of a virtual rodent. *arXiv preprint arXiv:1911.09451*, 2019.

[82] Eyrun Eyjolfsdottir, Kristin Branson, Yisong Yue, and Pietro Perona. Learning recurrent representations for hierarchical behavior modeling. *ICLR*, 2016.

[83] Jan Clemens and Mala Murthy. The use of computational modeling to link sensory processing with behavior in drosophila. In *Decoding Neural Circuit Structure and Function*, pages 241–260. Springer, 2017.

[84] Kiana Ehsani, Hessam Bagherinezhad, Joseph Redmon, Roozbeh Mottaghi, and Ali Farhadi. Who let the dogs out? modeling dog behavior from visual data. *CVPR*, 2018.

[85] Matthew M Botvinick. Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences*, 12(5):201–208, 2008.

[86] Mortimer Mishkin, Wendy A Suzuki, David G Gadian, and Faraneh Vargha-Khadem. Hierarchical organization of cognitive memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1360):1461–1467, 1997.

[87] Peter Meltzer and Peter J Bentley. Interacting hierarchical dynamic networks. In *Artificial Life Conference Proceedings*, pages 582–589. MIT Press, 2018.

[88] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

[89] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

[90] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[91] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135, 1999.

[92] Alec Solway, Carlos Diuk, Natalia Córdova, Debbie Yee, Andrew G Barto, Yael Niv, and Matthew M Botvinick. Optimal behavioral hierarchy. *PLoS computational biology*, 10(8):e1003779, 2014.

[93] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

[94] Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan Gecsei. On the optimization of a synaptic

learning rule. In *Preprints Conf. Optimality in Artificial and Biological Neural Networks*, volume 2. Univ. of Texas, 1992.

[95] Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):1–7, 2019.

[96] Gary Marcus. Innateness, alphazero, and artificial intelligence. *arXiv preprint arXiv:1801.05667*, 2018.

[97] Gordon J Berman, Daniel M Choi, William Bialek, and Joshua W Shaevitz. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface*, 11(99):20140672, 2014.

[98] Philip N Lehner. *Handbook of ethological methods*. Cambridge University Press, 1998.

[99] Steven M Frankland and Joshua D Greene. Concepts and compositionality: In search of the brain's language of thought. *Annual review of psychology*, 71, 2019.

[100] Raymond B Cattell. Abilities: Their structure, growth, and action. 1971.

[101] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[102] Adele Diamond. Executive functions. *Annual review of psychology*, 64:135–168, 2013.

[103] Eric Jonathan Nestler, Steven E Hyman, and Robert C Malenka. *Molecular neuropharmacology: a foundation for clinical neuroscience*. McGraw-Hill Medical, 2001.

[104] Jeffrey M Zacks and Khena M Swallow. Event segmentation. *Current directions in psychological science*, 16(2):80–84, 2007.

[105] Hoang M Le, Nan Jiang, Alekh Agarwal, Miroslav Dudík, Yisong Yue, and Hal Daumé III. Hierarchical imitation and reinforcement learning. *arXiv preprint arXiv:1803.00590*, 2018.

[106] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. *arXiv preprint arXiv:1709.10089*, 2017.

[107] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Gabriel Dulac-Arnold, et al. Deep q-learning from demonstrations. *arXiv preprint arXiv:1704.03732*, 2017.

[108] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[109] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[110] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[111] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[112] Christopher M Bishop. Mixture density networks. Technical report, Citeseer, 1994.

[113] Dhanushi A Wijeyakulasuriya, Elizabeth W Eisenhauer, Benjamin A Shaby, and Ephraim M Hanks. Machine learning for modeling animal movement. *Plos one*, 15(7):e0235750, 2020.

[114] Ella Browning, Mark Bolton, Ellie Owen, Akiko Shoji, Tim Guilford, and Robin Freeman. Predicting animal behaviour using deep learning: Gps data alone accurately predict diving in seabirds. *Methods in Ecology and Evolution*, 9(3):681–692, 2018.

[115] Zac Yung-Chun Liu, Jerry H Moxley, Paul Kanive, Adrian C Gleiss, Thom Maughan, Larry Bird, Oliver JD Jewell, Taylor K Chapple, Tyler Gagne, Connor F White, et al. Deep learning accurately predicts white shark locomotor activity from depth data. *Animal Biotelemetry*, 7(1):1–13, 2019.

[116] Daniel Gutierrez-Galan, Juan P Dominguez-Morales, Elena Cerezuela-Escudero, Antonio Rios-Navarro, Ricardo Tapiador-Morales, Manuel Rivas-Perez, Manuel Dominguez-Morales, Angel Jimenez-Fernandez, and Alejandro Linares-Barranco. Embedded neural network for real-time animal behavior classification. *Neurocomputing*, 272:17–26, 2018.

[117] Ahmet Arac, Pingping Zhao, Bruce H Dobkin, S Thomas Carmichael, and Peyman Golshani. Deepbehavior: A deep learning toolbox for automated analysis of animal and human behavior imaging data. *Frontiers in systems neuroscience*, 13:20, 2019.

[118] Talmo D Pereira, Diego E Aldarondo, Lindsay Willmore, Mikhail Kislin, Samuel S-H Wang, Mala Murthy, and Joshua W Shaevitz. Fast animal pose estimation using deep neural networks. *Nature methods*, 16(1):117–125, 2019.

[119] Alex Borowicz, Hieu Le, Grant Humphries, Georg Nehls, Caroline Höschle, Vladislav Kosarev, and

Heather J Lynch. Aerial-trained deep learning networks for surveying cetaceans from satellite imagery. *PloS one*, 14(10):e0212532, 2019.

[120] André C Ferreira, Liliana R Silva, Francesco Renna, Hanja B Brandl, Julien P Renoult, Damien R Farine, Rita Covas, and Claire Doutrelant. Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, 11(9):1072–1085, 2020.

[121] Melanie Clapham, Ed Miller, Mary Nguyen, and Chris T Darimont. Automated facial recognition for wildlife that lack unique markings: A deep learning approach for brown bears. *Ecology and evolution*, 10(23):12883–12892, 2020.

[122] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014.

[123] Ulrich Stern, Ruo He, and Chung-Hui Yang. Analyzing animal behavior via classifying each video frame using convolutional neural networks. *Scientific reports*, 5:14351, 2015.

[124] Sandeep Robert Datta, David J Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. Computational neuroethology: a call to action. *Neuron*, 104(1):11–24, 2019.

[125] Lila L Gatlin et al. Information theory and the living system. 1972.

[126] John Maynard Smith. The concept of information in biology. *Philosophy of science*, 67(2):177–194, 2000.

[127] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[128] F. Rieke, D. Warland, Rob. de Ruyter van, Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA, USA, 1999.

[129] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, USA, 2006.

[130] M. M. Ueli. The role of information theory in cryptography. In *Fourth IMA Conference on Cryptography and Coding*, pages 49–71, 1993.

[131] M. Nykter, N. D. Price, M. Aldana, N. A. Ramsey, S. A. Kauffman, L. E. Hood, O. Yli-Harja, and I. Shmulevich. Gene expression dynamics in the macrophage exhibit criticality. *Proceedings of*

the National Academy of Sciences, 105(6):1897–1900, 2008. doi: 10.1073/pnas.0711525105. URL http://www.pnas.org/content/105/6/1897.

[132] Macchiavello. *Quantum Computation and Quantum Information Theory*. World Scientific Publishing, Singapore, River Edge, N. J., 2000.

[133] Joseph T Lizier, Mikhail Prokopenko, and Albert Y Zomaya. A framework for the local information dynamics of distributed computation in complex systems. In *Guided self-organization: inception*, pages 115–158. Springer, 2014.

[134] Jennifer M Miller, X Rosalind Wang, Joseph T Lizier, Mikhail Prokopenko, and Louis F Rossi. Measuring information dynamics in swarms. In *Guided self-organization: Inception*, pages 343–364. Springer, 2014.

[135] A. J. King, A. M. Wilson, S. D. Wilshin, J. Lowe, H. Haddadi, S. Hailes, and A. J. Morton. Selfish-herd behaviour of sheep under threat. *Current Biology*, 22(14):R561–R562, 2012. doi: http://dx.doi.org/10.1016/j.cub.2012.05.008. URL http://www.cell.com/current-biology/abstract/S0960-9822(12)00529-5.

[136] M. Nagy, Z. Akos, D. Biro, and T. Vicsek. Hierarchical group dynamics in pigeon flocks. *Nature*, 464 (7290):890–893, 2010. doi: 10.1038/nature08891. URL https://www.nature.com/nature/journal/v464/n7290/full/nature08891.html.

[137] J. T. Mandel, K. L. Bildstein, G. Bohrer, and D. W. Winkler. Movement ecology of migration in turkey vultures. *Proceedings of the National Academy of Sciences*, 105(49):19102–19107, 2008. doi: 10.1073/pnas.0801789105. URL http://www.pnas.org/content/105/49/19102.

[138] M.K. Boyland, D.T. Mlynski, R. James, L. J. N. Brent, and D. P. Croft. The social network structure of a dynamic group of dairy cows: From individual to group level patterns. *Applied Animal Behaviour Science*, 174:1–10, 2016. doi: http://dx.doi.org/10.1016/j.applanim.2015.11.016. URL http://www.sciencedirect.com/science/article/pii/S0168159115003202.

[139] Ioannis Psorakis, Stephen J Roberts, Iead Rezek, and Ben C Sheldon. Inferring social network structure in ecological systems from spatio-temporal data streams. *Journal of the Royal Society Interface*, 9(76): 3055–3066, 2012.

[140] Daniele Neri, Tommaso Ruberto, Gabrielle Cord-Cruz, and Maurizio Porfiri. Information theory and

robotics meet to study predator-prey interactions. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(7):073111, 2017.

[141] Feng Hu, Li-Juan Nie, and Shi-Jian Fu. Information dynamics in the interaction between a prey and a predator fish. *Entropy*, 17(10):7230–7241, 2015.

[142] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27 (3):379–423, 1948.

[143] C.E. Shannon and W. Weaver. *The mathematical theory of communication.* University of Illinois Press, 1949.

[144] Peter Grunwald and Paul Vitányi. Shannon information and kolmogorov complexity. *arXiv preprint cs/0410002*, 2004.

[145] A. Rènyi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.

[146] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1): 479–487, 1988. doi: 10.1007/BF01016429. URL https://doi.org/10.1007/BF01016429.

[147] W. M. Lord, J. Sun, N. T. Ouellette, and E. M. Bolt. Inference of causal information flow in collective animal behavior. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2 (1):107–116, 2016. doi: 10.1109/TMBMC.2016.2632099. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7809221.

[148] J Alfred Rider, Dean L Rider, John M Opitz, James F Reynolds, and Raju K Pullarkat. Batten disease: past, present, and future. *American Journal of Medical Genetics Part A*, 31(S5):21–26, 1988.

[149] Terry J Lerner, Rose-Mary N Boustany, John W Anderson, Kenneth L D'Arigo, Karen Schlumpf, Alan J Buckler, James F Gusella, and Jonathan L Haines. Isolation of a novel gene underlying batten disease, CLN3. *Cell*, 82(6):949–957, 1995.

[150] T. Furmston, A. J. Morton, and S. Hailes. A significance test for inferring affiliation networks from spatio-temporal data. *PLOS ONE*, 10(7):1–23, 2015. doi: 10.1371/journal.pone.0132417. URL https://doi.org/10.1371/journal.pone.0132417.

[151] B. Jiang. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution.

The Professional Geographer, 65(3):482–494, 2013. doi: 10.1080/00330124.2012.700499. URL http://dx.doi.org/10.1080/00330124.2012.700499.

[152] D. J. T. Sumpter. *Collective Animal Behavior*. Princeton University Press, 2010.

[153] J. Krause, D. Lusseau, and R. James. Animal social networks: an introduction. *Behavioral Ecology and Sociobiology*, 63(7):967–973, 2009. doi: 10.1007/s00265-009-0747-0. URL http://dx.doi.org/10.1007/s00265-009-0747-0.

[154] D. W. Franks, G. D. Ruxton, and R. James. Sampling animal association networks with the gambit of the group. *Behavioral Ecology and Sociobiology*, 64(3):493–503, 2010. doi: 10.1007/s00265-009-0865-8. URL http://dx.doi.org/10.1007/s00265-009-0865-8.

[155] D. R. Farine. A guide to null models for animal social network analysis. *Methods in Ecology and Evolution*, 8(10):1309–1320, 2017. doi: 10.1111/2041-210X.12772. URL http://dx.doi.org/10.1111/2041-210X.12772.

[156] C. D. Santos, S. Neupert, H-P. Lipp, M. Wikeski, and D. K. N. Dechmann. Temporal and contextual consistency of leadership in homing pigeon flocks. *PLoS One*, 9(7):e102771, 2014. doi: 10.1371/journal.pone.0102771. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4108361/.

[157] L. Van Bommel and C. N. Johnson. Data from: Where do livestock guardian dogs go? movement patterns of free-ranging maremma sheepdogs. *Movebank data repository*, 2014. doi: 10.5441/001/1.pv048q7v.

[158] Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

[159] Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. In *Advances in Neural Information Processing Systems*, pages 5986–5997, 2017.

[160] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[161] Joseph T Lizier. Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1:11, 2014.

[162] P. M. B. Vitanyi, F. J. Balbach, , R. L. Cilibrasi, and M. Li. *Information Theory and Statistical Learning:Normalized Information Distance*. Springer US, Boston, MA, 2009.

[163] R. Cilibrasi and P. Vitanyi. Clustering by compression. *IEEE Trans. Inf. Theor.*, 51(4):1523–1545, 2005. doi: 10.1109/TIT.2005.844059. URL http://dx.doi.org/10.1109/TIT.2005.844059.

[164] E. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards parameter-free data mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 206–215. ACM, 2004.

[165] Z. Cataltepe, Y. Yaslan, and A. Sonmez. Music genre classification using midi and audio features. *EURASIP Journal on Advances in Signal Processing*, 2007(1):036409, 2007. doi: 10.1155/2007/36409. URL https://doi.org/10.1155/2007/36409.

[166] S. Wehner. Analyzing worms and network traffic using compression. *Journal of Computer Security*, 15(3):303–320, 2007. doi: 10.3233/JCS-2007-15301. URL http://dl.acm.org/citation.cfm?id=1370628.1370630.

[167] Wayne Hsiung and Cass R Sunstein. Climate change and animals. *University of Pennsylvania Law Review.*, 155:1695, 2006.

[168] B. Abrahms, D. P. Seidel, E. Dougherty, E.L. Hazen, S. J. Bograd, A. M. Wilson, J. W. McNutt, Costa D. P., S. Blake, J. S. Brashares, and Getz W. M. Suite of simple metrics reveals common movement syndromes across vertebrate taxa. *Movement Ecology*, 5(1):12, 2017. doi: 10.1186/s40462-017-0104-2. URL https://doi.org/10.1186/s40462-017-0104-2.

[169] S. Blake, C. B. Yackulic, F. Cabrera, W. Tapia, J.P. Gibbs, F. Kummeth, and M. Wilkeski. Vegetation dynamics drive segregation by body size in galapagos tortoises migrating across altitudinal gradients. *Journal of Animal Ecology*, 82(2):310–321, 2013. doi: 10.1111/1365-2656.12020. URL http://dx.doi.org/10.1111/1365-2656.12020.

[170] B. Abrahms. Data from: Suite of simple metrics reveals common movement syndromes across vertebrate taxa. movebank data repository. *Movebank data repository*, 2017. doi: 10.5441/001/1.hm5nk220.

[171] J. L. Andrews, C. T. Wendy, and M. G. Wayne. Home range plus: a space-time characterization of movement over real landscapes. *Movement Ecology*, 1(1):2, 2013. doi: 10.1186/2051-3933-1-2. URL https://doi.org/10.1186/2051-3933-1-2.

[172] S. Bar-David, I. Bar-David, P. C. Cross, S. J. Ryan, C. U. Knechtel, and W. M. Getz. Methods for

assessing movement path recursion with application to african buffalo in south africa. *Ecology*, 90(9): 2467–2479, 2009. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3025599/.

[173] P. C. Cross, J. A. Bowers, C. T. Hay, J. Wolhuter, P. Buss, M. Hofmeyr, J. T. du Toit, and W. M. Getz. Data from: Nonparameteric kernel methods for constructing home ranges and utilization distributions. *Movebank data repository*, 2016. doi: 10.5441/001/1.j900f88t.

[174] W. M. Getz, S. Fortmann-Roe, P. C. Cross, A. J. Lyons, S. J. Ryan, and C. C. Wilmers. LoCoH: Nonparameteric kernel methods for constructing home ranges and utilization distributions. *PLoS ONE*, 2(2), 2007. doi: doi:10.1371/journal.pone.0000207. URL https://doi.org/10.1371/journal.pone.0000207.

[175] S. E. Bellan, C. A. Cizauskas, J. Miyen, K. Ebersohn, M. Kusters, K. Prager, M. Van Vuuren, C. Sabeta, and W. M. Getz. Black-backed jackal exposure to rabies virus, canine distemper virus and bacillus anthracis in Etosha national park, Namibia. *Journal of wildlife diseases*, 48(2):371–381, 2012. doi: 10.7589/0090-3558-48.2.371. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5479413/.

[176] S. Dodge, G. Bohrer, R. Weinzierl, S. C. Davidson, R. Kays, D. Douglas, S. Cruz, J. Han, D. Brandes, and M. Wikelski. The environmental-data automated track annotation (env-data) system: Linking animal tracks with environmental data. *Movement Ecology*, 1(1):3, 2013. doi: 10.1186/2051-3933-1-3. URL https://doi.org/10.1186/2051-3933-1-3.

[177] L. Van Bommel and C. N. Johnson. Where do livestock guardian dogs go? movement patterns of free-ranging maremma sheepdogs. *PLoS ONE*, 9(10):e111444, 2014. doi: 10.1371/journal.pone.0111444. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4213034/.

[178] P. W. Robinson, D. P. Costa, D. E. Crocker, J. P. Gallo-Reynoso, C. D. Champagne, M. A. Fowler, C. Goetsch, K. T. Goetz, J. L. Hassrick, C. E. Kuhn, J. L. Maresh, S. M. Maxwell, B. I. McDonald, S. H. Peterson, S. E. Simmons, N. M. Teutschel, S. Villegas-Amtmann, and K. Yoda. Foraging behavior and success of a mesopelagic predator in the northeast pacific ocean: Insights from a data-rich species, the northern elephant seal. *PLoS ONE*, 7(5):e36728, 2012. doi: 10.1371/journal.pone.0036728. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3352920/.

[179] O. Spiegel, W. M. Getz, and R. Nathan. Factors influencing foraging search efficiency: Why do scarce lappet-faced vultures outperform ubiquitous white-backed vultures? *The American Naturalist*, 181(5): E102–E115, 2013. doi: 10.1086/655823. URL http://www.jstor.org/stable/10.1086/670009.

[180] O. Spiegel, W. M. Getz, and R. Nathan. Data from: Factors influencing foraging search efficiency: Why do scarce lappet-faced vultures outperform ubiquitous white-backed vultures? *Movebank data repository*, 2014. doi: 10.5441/001/1.mf903197.

[181] H. L. A. Bartlam-Brooks, P. S. A. Beck, G. Bohrer, and S. Harris. In search of greener pastures: Using satellite images to predict the effects of environmental change on zebra migration? *Journal of Geophysical Research: Biogeosciences*, 118(4):1427–1437, 2013. doi: 10.1002/jgrg.20096. URL http://dx.doi.org/10.1002/jgrg.20096.

[182] H. L. A. Bartlam-Brooks and S. Harris. Data from: In search of greener pastures: using satellite images to predict the effects of environmental change on zebra migration. *Movebank data repository*, 2013. doi: 10.5441/001/1.f3550b4f.

[183] Pablo Montero and José A. Vilar. TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1):1–43, 2014. URL http://www.jstatsoft.org/v62/i01/.

[184] M. R. Hirt, W. Jetz, B. C. Rall, and U. Brose. A general scaling law reveals why the largest animals are not the fastest. *Nature Ecology & Evolution*, 1(8):1116–1122, 2017. doi: 10.1038/s41559-017-0241-4. URL https://doi.org/10.1038/s41559-017-0241-4.

[185] G. A. Breed, P. M. Severns, and A. M. Edwards. Apparent power-law distributions in animal movements can arise from intraspecific interactions. *Journal of the Royal Society Interface*, 12(103):1116–1122, 2015. doi: 10.1098/rsif.2014.0927. URL http://doi.org/10.1098/rsif.2014.0927.

[186] Donald F Caldwell and Edward F Domino. Electroencephalographic and eye movement patterns during sleep in chronic schizophrenic patients. *Electroencephalography and Clinical Neurophysiology*, 22(5): 414–420, 1967.

[187] Vijay A Mittal and Elaine F Walker. Dyskinesias, tics, and psychosis: Issues for the next diagnostic and statistical manuel of mental disorders. *Psychiatry research*, 189(1):158, 2011.

[188] RA Blatchford, KC Klasing, HL Shivaprasad, PS Wakenell, GS Archer, and JA Mench. The effect of light intensity on the behavior, eye and leg health, and immune function of broiler chickens. *Poultry science*, 88(1):20–28, 2009.

[189] Frank Miskelly. A novel system of electronic tagging in patients with dementia and wandering. *Age and ageing*, 33(3):304–306, 2004.

[190] Jenny Louise Gibson, Stephen Hailes, Behzad Heravi, and David Skuse. Using sensors to study the social dynamics of outdoor play. 2018.

[191] Gilles Allali, Helena M Blumen, Hervé Devanne, Elvira Pirondini, Arnaud Delval, and Dimitri Van De Ville. Brain imaging of locomotion in neurological conditions. *Neurophysiologie Clinique*, 48(6): 337–359, 2018.

[192] Heather R Adams, Jennifer Kwon, Frederick J Marshall, Elisabeth A de Blieck, David A Pearce, and Jonathan W Mink. Neuropsychological symptoms of juvenile-onset batten disease: experiences from 2 studies. *Journal of child neurology*, 22(5):621–627, 2007.

[193] Sarah-Bianca Dolisca, Mitali Mehta, David A Pearce, Jonathan W Mink, and Bernard L Maria. Batten disease: clinical aspects, molecular mechanisms, translational science, and future directions. *Journal of child neurology*, 28(9):1074–1100, 2013.

[194] Patricia Gilbert. *The AZ reference book of syndromes and inherited disorders*. Springer, 2013.

[195] Kim Wager, Anselm A Zdebik, Sonia Fu, Jonathan D Cooper, Robert J Harvey, and Claire Russell. Neurodegeneration and epilepsy in a zebrafish model of cln3 disease (batten disease). *PloS one*, 11(6): e0157365, 2016.

[196] Fabiola Mara Ribeiro, Elizabeth Ribeiro da Silva Camargos, Leonardo Cruz de Souza, and Antonio Lucio Teixeira. Animal models of neurodegenerative diseases. *Revista brasileira de psiquiatria*, 35:S82–S91, 2013.

[197] Cathleen M Lutz and Melissa A Osborne. Optimizing mouse models of neurodegenerative disorders: are therapeutics in sight? *Future Neurology*, 9(1):67–75, 2014.

[198] Debra J Skene, Benita Middleton, Cara K Fraser, Jeroen LA Pennings, Timothy R Kuchel, Skye R Rudiger, C Simon Bawden, and A Jennifer Morton. Metabolic profiling of presymptomatic huntington?s disease sheep reveals novel biomarkers. *Scientific Reports*, 7:43030, 2017.

[199] RD Jolly, DG Arthur, GW Kay, and DN Palmer. Neuronal ceroid-lipofuscinosis in borderdale sheep. *New Zealand veterinary journal*, 50(5):199–202, 2002.

[200] Hamed Haddadi, Andrew J King, Alison P Wills, Damien Fay, John Lowe, A Jennifer Morton, Stephen Hailes, and Alan M Wilson. Determining association networks in social animals: choosing

spatial–temporal criteria and sampling rates. *Behavioral Ecology and Sociobiology*, 65(8):1659–1668, 2011.

[201] Hannah Hobbs-Chell, Andrew J King, Hannah Sharratt, Hamed Haddadi, Skye R Rudiger, Stephen Hailes, A Jennifer Morton, and Alan M Wilson. Data-loggers carried on a harness do not adversely affect sheep locomotion. *Research in veterinary science*, 93(1):549–552, 2012.

[202] T Takasu. Rtklib: An open source program package for gnss positioning. *Tech. Rep., 2013. Software and documentation*, 2011.

[203] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[204] Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. Outlier detection using k-nearest neighbour graph. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 430–433. IEEE, 2004.

[205] Shruti Aggarwal and Janpreet Singh. Outlier detection using k-mean and hybrid distance technique on multi-dimensional data set. *International Journal of Advanced Research in Computer Engineering and Technology*, 2(9):2626–31, 2013.

[206] Dan Pelleg, Andrew W Moore, et al. X-means: extending k-means with efficient estimation of the number of clusters. In *Icml*, volume 1, pages 727–734, 2000.

[207] Kehinde Owoeye and Stephen Hailes. Online collective animal movement activity recognition. *Modeling and decision-making in the spatiotemporal domain workshop, NeurIPS*, 2018.

[208] G. A. United Nations. Transforming our world: the 2030 agenda for sustainable development. New York: United Nations, 2015.

[209] Çağan H Şekercioğlu, Gretchen C Daily, and Paul R Ehrlich. Ecosystem consequences of bird declines. *Proceedings of the National Academy of Sciences*, 101(52):18042–18047, 2004.

[210] Kenneth V Rosenberg, Adriaan M Dokter, Peter J Blancher, John R Sauer, Adam C Smith, Paul A Smith, Jessica C Stanton, Arvind Panjabi, Laura Helft, Michael Parr, et al. Decline of the north american avifauna. *Science*, 366(6461):120–124, 2019.

[211] Hugh A Ford, David C Paton, and Neville Forde. Birds as pollinators of australian plants. *New Zealand journal of botany*, 17(4):509–519, 1979.

[212] Steven F Railsback and Matthew D Johnson. Effects of land use on bird populations and pest control services on coffee farms. *Proceedings of the National Academy of Sciences*, 111(16):6109–6114, 2014.

[213] Lajos Rozsa. Patterns in the abundance of avian lice (phthiraptera: Amblycera, ischnocera). *Journal of Avian Biology*, pages 249–254, 1997.

[214] Scott R Loss, Tom Will, Sara S Loss, and Peter P Marra. Bird–building collisions in the united states: Estimates of annual mortality and species vulnerability. *The Condor*, 116(1):8–23, 2014.

[215] Scott R Loss, Tom Will, and Peter P Marra. Refining estimates of bird collision and electrocution mortality at power lines in the united states. *PLoS One*, 9(7):e101565, 2014.

[216] Benjamin M Van Doren, Kyle G Horton, Adriaan M Dokter, Holger Klinck, Susan B Elbin, and Andrew Farnsworth. High-intensity urban light installation dramatically alters nocturnal bird migration. *Proceedings of the National Academy of Sciences*, 114(42):11175–11180, 2017.

[217] Somayeh Dodge, Gil Bohrer, Keith Bildstein, Sarah C Davidson, Rolf Weinzierl, Marc J Bechard, David Barber, Roland Kays, David Brandes, Jiawei Han, et al. Environmental drivers of variability in the movement ecology of turkey vultures (cathartes aura) in north and south america. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1643):20130195, 2014.

[218] Peter Berthold. *Bird migration: a general survey*. Oxford University Press on Demand, 2001.

[219] Friedemann Koester. Observations on migratory turkey vultures and lesser yellow-headed vultures in northern colombia. *The Auk*, pages 372–375, 1982.

[220] Maricel Graña Grilli, Sergio A Lambertucci, Jean-François Therrien, and Keith L Bildstein. Wing size but not wing shape is related to migratory behavior in a soaring bird. *Journal of avian biology*, 48(5): 669–678, 2017.

[221] Keith L Bildstein, David Barber, Marc J Bechard, and Maricel Grana Grilli. Data from: Wing size but not wing shape is related to migratory behavior in a soaring bird. 2016.

[222] Somayeh Dodge, Gil Bohrer, Rolf Weinzierl, Sarah C Davidson, Roland Kays, David Douglas, Sebastian Cruz, Jiawei Han, David Brandes, and Martin Wikelski. The environmental-data automated track annotation (env-data) system: linking animal tracks with environmental data. *Movement Ecology*, 1(1): 3, 2013.

[223] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger

Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[224] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[225] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, pages 6775–6785, 2018.

[226] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *International Conference on Learning Representations*, 2017.

[227] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[228] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699. ACM, 2002.

[229] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

[230] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[231] Nathan D Wolfe, Claire Panosian Dunavan, and Jared Diamond. Origins of major human infectious diseases. *Nature*, 447(7142):279–283, 2007.

[232] Anders Hedenström and Thomas Alerstam. Optimum fuel loads in migratory birds: distinguishing between time and energy minimization. *Journal of theoretical biology*, 189(3):227–234, 1997.

[233] Kasper Thorup, Anders P Tøttrup, Mikkel Willemoes, Raymond HG Klaassen, Roine Strandberg, Marta Lomas Vega, Hari P Dasari, Miguel B Araújo, Martin Wikelski, and Carsten Rahbek. Resource tracking within and across continents in long-distance bird migrants. *Science Advances*, 3(1):e1601360, 2017.

[234] Sotirios Tsiodras, Theodoros Kelesidis, Iosif Kelesidis, Ulf Bauchinger, and Matthew E Falagas. Human infections associated with wild birds. *Journal of Infection*, 56(2):83–98, 2008.

[235] Ruth E Crawford and Jed A Long. Habitat preferences of juvenile scottish ospreys pandion haliaetus at stopover and wintering sites. *Ringing & Migration*, 32(1):1–18, 2017.

[236] Vanessa Brum-Bastos, Jed Long, Katharyn Church, Greg Robson, Rogério de Paula, and Urška Demšar. Multi-source data fusion of optical satellite imagery to characterize habitat selection from wildlife tracking data. *Ecological Informatics*, 60:101149, 2020.

[237] Kehinde Owoeye. Preventing future outbreaks: A case for global surveillance of wild migratory birds.

[238] Andrea Kölzsch, Gerhard JDM Müskens, Helmut Kruckenberg, Peter Glazov, Rolf Weinzierl, Bart A Nolet, and Martin Wikelski. Towards a new understanding of migration timing: slower spring than autumn migration in geese reflects different decision rules for stopover use and departure. *Oikos*, 125 (10):1496–1507, 2016.

[239] A Kölzsch, H Kruckenberg, P Glazov, GJDM Müskens, and M Wikelski. Data from: Towards a new understanding of migration timing: slower spring than autumn migration in geese reflects different decision rules for stopover use and departure, 2016. URL http://dx.doi.org/10.5441/001/1.31c2v92f.

[240] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[241] Orlando Gomez, Rafael López, and Adrián Naveda-Rodriguez. Avian pox in turkey vulture (cathartes aura). *Vulture News*, 66:42–48, 2014.

[242] Svein-Håkon Lorentsen, Ingar Jostein Øien, and Tomas Aarvak. Migration of fennoscandian lesser white-fronted geese anser erythropus mapped by satellite telemetry. *Biological Conservation*, 84(1): 47–52, 1998.

[243] F Reed Hainsworth. Precision and dynamics of positioning by canada geese flying in formation. *Journal of Experimental Biology*, 128(1):445–462, 1987.

[244] Johan Elmberg, Charlotte Berg, Henrik Lerner, Jonas Waldenström, and Rebecca Hessel. Potential disease transmission from wild geese and swans to livestock, poultry and humans: a review of the scientific literature from a one health perspective. *Infection ecology & epidemiology*, 7(1):1300450, 2017.

[245] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019.

[246] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[247] Charles Fergus. *Wildlife of Virginia and Maryland and Washington, DC*. Stackpole Books, 2003.

[248] Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 2020.

[249] Franziska Knolle, Rita P Goncalves, and A Jennifer Morton. Sheep recognize familiar and unfamiliar human faces from two-dimensional images. *Royal Society open science*, 4(11):171228, 2017.

[250] Keith M Kendrick, Ana P da Costa, Andrea E Leigh, Michael R Hinton, and Jon W Peirce. Sheep don't forget a face. *Nature*, 414(6860):165, 2001.

[251] Alex J Champandard and Philip Dunstan. The behavior tree starter kit. In *Game AI Pro 360*, pages 27–46. CRC Press, 2019.

[252] Tobias Glasmachers. Limits of end-to-end learning. *Proceedings of Machine Learning Research, ACML*, 2017.

[253] John Haugeland. *Artificial intelligence: The very idea*. MIT press, 1989.

[254] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015.

[255] Panna Felsen, Patrick Lucey, and Sujoy Ganguly. Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 732–747, 2018.

[256] Craig W Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 25–34, 1987.

[257] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009.

[258] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[259] Kehinde Owoeye, Mirco Musolesi, and Stephen Hailes. Quantifying unusual neurological movement phenotypes in collective movement phenotypes. *bioRxiv*, 2021.

[260] Kehinde Owoeye. Forecasting avian migration patterns using a deep bidirectional rnn augmented with an auxiliary task. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4382–4388, 2020.

[261] Kehinde Owoeye. Deep rnn with pseudo loss objective for forecasting stop-over decisions of wild migratory birds. *bioRxiv*, 2021.

Appendix A

Characterizing Animal Movement Patterns Using Information Theory (Supplementary Material)

All computations were made using MATLAB R2017a except for the Normalized Compression Distance that was computed using the TSclust package in R and the mutual information for discrete and continuous mixtures in python.

A.0.1 Case Study: Shannon entropy as a tool for characterizing movement patterns of sheep with neurodegenerative disease

For the computations, I used 12 bins (symbols) to compute the entropy. Distance covered every ten minutes was used to generate the symbols. To bin the data, I used the head/tail classification rule by [151] where the data was divided mainly in the direction of skewness around the mean until there was only one data point left.

A.0.2 Mutual information for measuring association in pigeons

I computed the turn-angle between contiguous data points for each bird using:

$$turn - angle = \text{atan2}\left(\frac{y_2 - y_1}{x_2 - x_1}\right) \tag{A.1}$$

Essentially, I computed the tan inverse of the ratio of the change in the x and y coordinates i.e. $(\frac{\delta y}{\delta x})$ using

contiguous points and subsequently computed the difference between contiguous turn angles. I then used the time series of this turn-angle to compute the pairwise mutual information across all birds in the flock.

A.0.2.1 Randomization Tests

To generate random data and learn significant mutual information, I used the code in Matlab: random data = min(min(data)) + (max(max(data))-min(min(data))).*(rand(size(data,1),size(data,2))). I computed the pairwise mutual information using the random data and repeat the experiment ten times. At each iteration I selected the maximum mutual information and averaged them after all iterations. I applied the same method when computing transfer entropy as well.

A.0.3 Case study: Kolmogorov complexity as tool for classifying animal movement patterns across scales

Here, I converted all decimals into binary, i.e., a series of zeroes and ones and aggregated all together over the period of observation for each animal. These strings of ones and zeroes in a pairwise manner were used in the following formula:

$$NCD(x,y) = \frac{Z(xy) - min\ \{Z(x), Z(y)\}}{max\ \{Z(x), Z(y)\}} \tag{A.2}$$

where the compressor Z (gzip) was used. I used the package 'TSclust' in R[1].

I compared the monthly movement patterns of all the 85 animals to find similarities by computing their pairwise NCD with the default compressor gzip followed by hierarchical clustering of the resulting distance matrix using the `TSclust` package in R.

The distance covered every one hour was transformed into its binary equivalent (strings of zeroes and ones) using the `as.numeric` float to binary conversion function in R. I also compared the annual movement patterns of 16 animals across 6 species using the same approach described above for the 85 animals.

[1] https://cran.r-project.org/web/packages/TSclust/TSclust.pdf

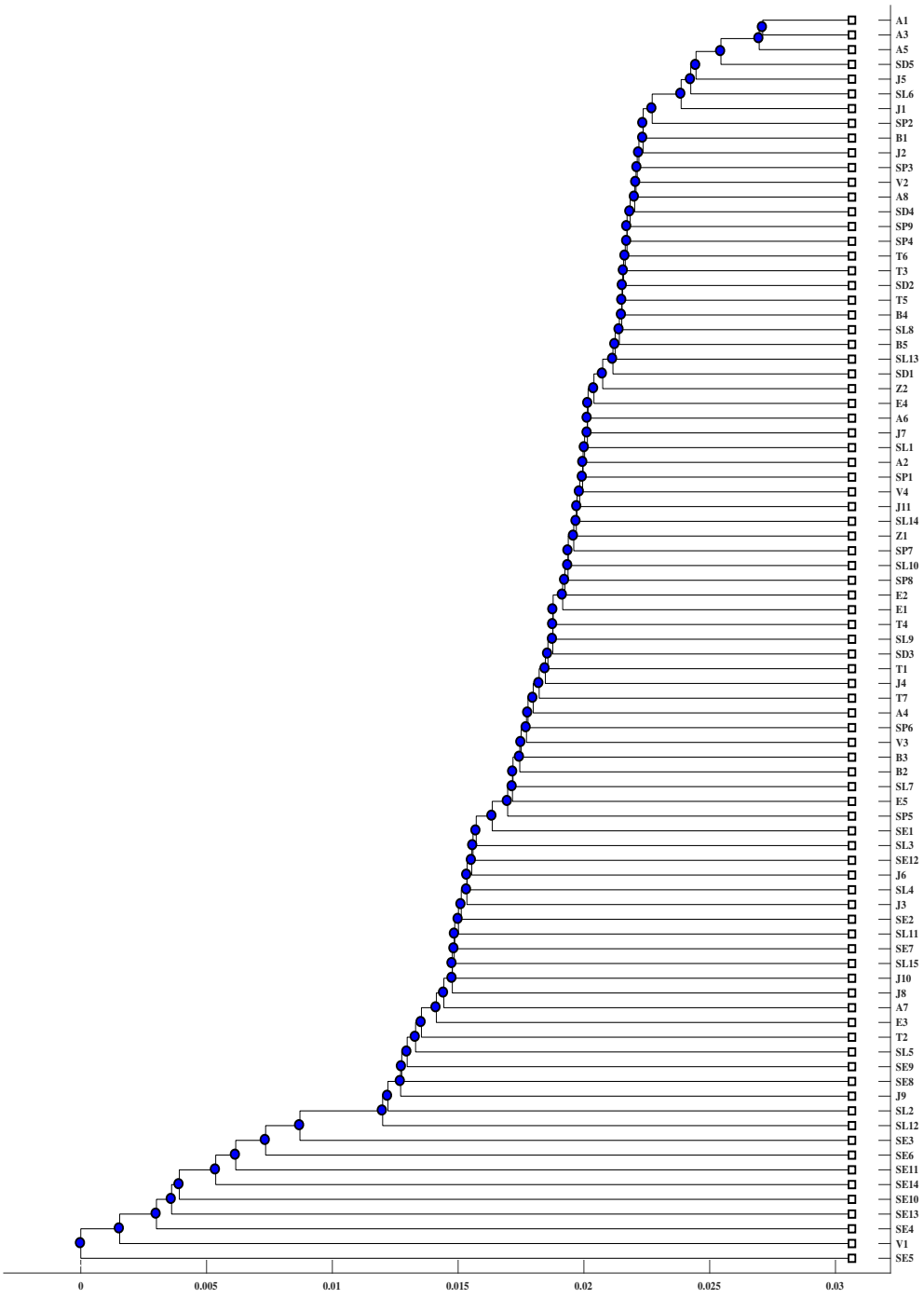


Figure A.1: **Hierarchical clustering.** Of the pairwise mutual information of the distance covered of 85 animals spread across 11 species representing the movement patterns over a period of one month.

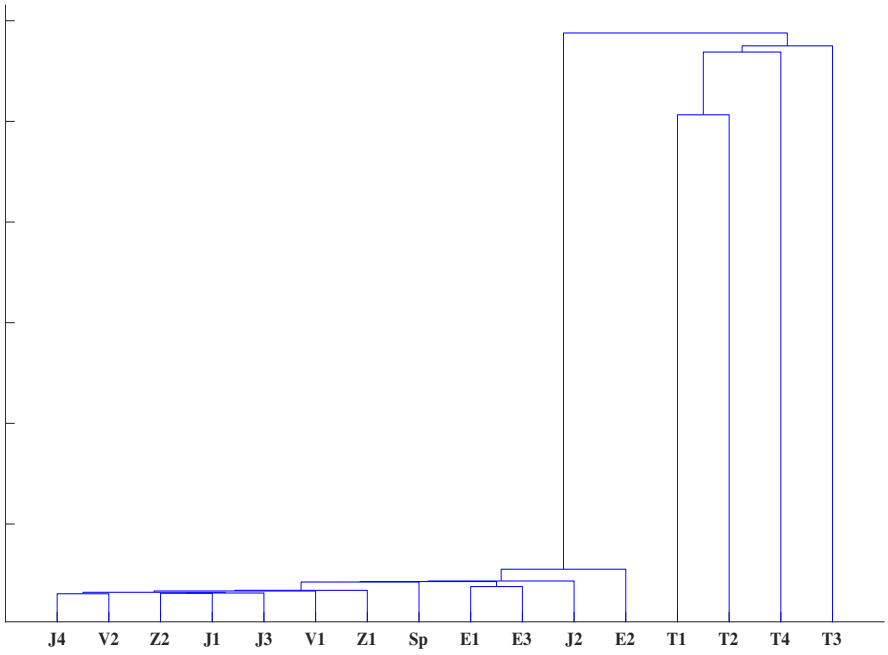


Figure A.2: **Hierarchical clustering.** Of the pairwise mutual information of the distance covered of 16 animals spread across 6 species representing the movement patterns over a period of one year.

A.1 Supplementary Tables and Figures Appendix A

Sheep ID	Batten Disease	Day1	Day2	Day3	Day4	Day5	Day6
1185.2009	YES	1.7645	1.3893	0.7607	1.6788	1.8132	1.8934
1187.2009	YES	1.8472	1.4395	1.9331	2.3363	1.9565	2.0052
1120.2009	YES	1.8962	2.3550	1.9297	1.7969	2.0628	2.334
1123.2009	YES	1.9823	1.7283	1.9629	0.1959	1.7149	1.8610
1156.2009	YES	2.0957	1.8789	1.3843	1.3276	1.4582	1.4025
1105.2009	NO	1.8432	2.0031	1.4586	2.6491	2.6392	2.1216
1169.2009	NO	2.3606	1.9889	1.8599	2.6295	1.9189	2.5715
1157.2009	NO	2.4714	2.5085	2.1798	2.6812	2.2815	2.7020
1181.2009	NO	2.4847	1.6326	1.6634	2.9731	2.2835	2.8444
1158.2009	NO	2.5434	2.4597	1.9509	2.5304	2.3482	2.0952
1193.2009	NO	2.6576	2.3378	2.2627	2.3872	1.9063	2.3367

Table A.1: **Entropy of 11 sheep in the flock across 6 days of observation.** A quick look at the table shows the batten sheep have a lower entropy relative to the normal sheep.

Pigeon ID	M	N	O	P	Q	R	S	T	U
M	-	0.0039	0.0039	0.0050	0.0023	0.0025	0.0101	0.0052	0.0037
N	0.0039	-	0.0357	0.0106	0.0069	0.0136	0.0172	0.0237	0.0129
O	0.0039	0.0357	-	0.0195	0.0199	0.0101	0.0596	0.0302	0.0123
P	0.0050	0.0106	0.0195	-	0.0240	0.0497	0.0178	0.0658	0.0167
Q	0.0023	0.0069	0.0199	0.0240	-	0.0428	0.0230	0.0295	0.0232
R	0.0025	0.0136	0.0101	0.0497	0.0428	-	0.0170	0.0591	0.0137
S	0.0101	0.0172	0.0596	0.0178	0.0230	0.0170	-	0.0264	0.0102
T	0.0052	0.0237	0.0302	0.0658	0.0295	0.0591	0.0264	-	0.0174
U	0.0037	0.0129	0.0123	0.0167	0.0232	0.0137	0.0102	0.0174	-

Table A.2: **Pairwise Mutual Information (Kernel Estimator) between the nine pigeons involved in the flight.**

Pigeon ID	M	N	O	P	Q	R	S	T	U
M	-	0.0062	0.0332	0.0169	0.0442	0.0276	0.0181	0.0428	0.0362
N	0.0062	-	0.1041	0.0771	0.0661	0.0626	0.0893	0.0866	0.1008
O	0.0332	0.1041	-	0.0820	0.1022	0.0952	0.1154	0.1028	0.0639
P	0.0169	0.0771	0.0820	-	0.0844	0.1155	0.0814	0.1037	0.1006
Q	0.0442	0.0661	0.1022	0.0844	-	0.1227	0.0735	0.0812	0.0507
R	0.0276	0.0626	0.0952	0.1155	0.1227	-	0.1143	0.1292	0.0529
S	0.0181	0.0893	0.1154	0.0814	0.0735	0.1143	-	0.0879	0.0839
T	0.0428	0.0866	0.1028	0.1037	0.0812	0.1292	0.0879	-	0.0973
U	0.0362	0.1008	0.0639	0.1006	0.0507	0.05293	0.0839	0.0973	-

Table A.3: **Pairwise Mutual Information (Discrete & continuous mixtures) between the nine pigeons involved in the flight.**

Pigeon ID	M	N	O	P	Q	R	S	T	U
M	-	-0.00248	0.0486	0.0292	0.06404	0.0230	0.0093	0.0438	0.0421
N	-0.0037	-	0.0997	0.0822	0.0733	0.0563	0.0865	0.0909	0.1052
O	0.0475	0.1013	-	0.0810	0.1182	0.0779	0.1135	0.1103	0.0559
P	0.0292	0.08233	0.0801	-	0.0752	0.1183	0.0848	0.1011	0.1004
Q	0.0644	0.0743	0.1193	0.0752	-	0.1063	0.0659	0.0765	0.0507
R	0.0227	0.0557	0.0770	0.1188	0.1051	-	0.1160	0.1282	0.0459
S	0.0095	0.0875	0.1136	0.0841	0.0647	0.1154	-	0.08973	0.0918
T	0.0435	0.0908	0.1102	0.1004	0.0770	0.1265	0.0906	-	0.0956
U	0.0431	0.1063	0.0545	0.1007	0.0509	0.0454	0.0921	0.0965	-

Table A.4: **Pairwise Mutual Information (Kraskov Estimator) between the nine pigeons involved in the flight.**

Pigeon ID	M	N	O	P	Q	R	S	T	U
M	-	0.0041	0.0037	0.0036	0.0091	0.0044	0.0026	0.0028	0.0067
N	0.0037	-	0.006	0.0136	0.0088	0.0075	0.0106	0.0081	0.0087
O	0.0049	0.0061	-	0.0107	0.0122	0.0066	0.0159	0.0054	0.0153
P	0.0056	0.0102	0.0163	-	0.0105	0.0076	0.0184	0.0094	0.0147
Q	0.0021	0.0143	0.0078	0.0039	-	0.0089	0.0102	0.0045	0.0127
R	0.0025	0.0074	0.0216	0.0153	0.0105	-	0.0158	0.0114	0.0125
S	0.0015	0.01	0.0167	0.0105	0.0073	0.0074	-	0.0131	0.0211
T	0.0023	0.0119	0.0154	0.012	0.0122	0.0093	0.0144	-	0.0161
U	0.0025	0.0059	0.0136	0.0113	0.0058	0.0085	0.005	0.0052	-

Table A.5: **Pairwise transfer entropy between the nine pigeons involved in the flight using a kernel estimator.**

Pigeon ID	M	N	O	P	Q	R	S	T	U
M	-	0.0142	0.015	0.0063	0.0264	0.0202	0.0167	0.0086	0.0177
N	0.0112	-	0.0458	0.0401	0.0517	0.0535	0.0166	0.0369	0.0392
O	0.0011	0.0803	-	0.073	0.042	0.078	0.0533	0.054	0.066
P	0.0175	0.0329	0.0675	-	0.0603	0.0569	0.0458	0.0485	0.0491
Q	0.0078	0.0577	0.0459	0.0645	-	0.0594	0.0944	0.0544	0.0653
R	0.0164	0.0464	0.0587	0.063	0.0625	-	0.0796	0.0409	0.0735
S	0.0125	0.0587	0.0475	0.0844	0.0618	0.0829	-	0.0475	0.0564
T	0.0155	0.0737	0.0729	0.069	0.0663	0.0479	0.0833	-	0.0631
U	0.0124	0.0418	0.0745	0.0561	0.0556	0.073	0.0617	0.0299	-

Table A.6: **Pairwise transfer entropy between the nine pigeons involved in the flight using a Kraskov estimator.**

Pigeon ID	M	N	O	P	Q	R	S	T	U
M	-	0.0244	-0.0140	-0.036	-0.013	-0.0021	0.0835	0.0114	-0.0571
N	0.0244	-	0.1882	0.0332	-0.0364	0.0744	0.0847	0.1366	0.0731
O	-0.0140	0.1882	-	0.0923	0.1108	0.0367	0.2584	0.1453	-0.0012
P	-0.036	0.0332	0.0923	-	0.1206	0.2379	0.067	0.2914	0.0899
Q	-0.013	-0.0364	0.1108	0.1206	-	0.2218	0.118	0.1774	0.145
R	-0.0021	0.00744	0.0367	0.2379	0.2218	-	0.0336	0.2671	0.0757
S	0.0835	0.0847	0.2584	0.067	0.118	0.0336	-	0.1133	-0.00029
T	0.0114	0.1366	0.1453	0.2914	0.1774	0.2671	0.1133	-	0.1026
U	-0.0571	0.0731	-0.0012	0.0899	0.145	0.0757	-0.00029	0.1026	-

Table A.7: **Pairwise correlation coefficient between the nine pigeons involved in the flight.**

Species ID	T1	T2	T3	T4	Z1	Z2	E1	E2	E3	J1	J2	J3	J4	Sp	V1	V2
T1	0.1034	0.9335	0.9361	0.9332	1.0231	1.0284	1.0297	1.0312	1.0349	1.0491	1.0490	1.0439	1.0447	1.0208	1.0751	1.0637
T2	0.9339	0.1087	0.9369	0.9223	1.0263	1.0262	1.0306	1.0308	1.0362	1.0568	1.0533	1.0568	1.0536	1.0297	1.0757	1.0552
T3	0.9360	0.9371	0.093	0.9344	1.0210	1.0232	1.0254	1.0266	1.0296	1.0488	1.0390	1.0513	1.0465	1.0254	1.0789	1.0666
T4	0.9334	0.9257	0.9339	0.1006	1.0272	1.0226	1.0294	1.0302	1.0321	1.0541	1.0521	1.0549	1.0518	1.0285	1.0793	1.0691
Z1	1.0205	1.0259	1.0162	1.0228	0.0037	0.9424	0.9488	0.9502	0.9535	1.0436	1.0189	1.0205	1.0377	0.9477	1.0325	1.0098
Z2	1.0244	1.0311	1.0198	1.0268	0.9467	0.0037	0.9509	0.9496	0.9545	1.0431	1.0409	1.0420	1.0266	0.9499	1.0327	1.0085
E1	1.0219	1.0287	1.0178	1.0251	0.9387	0.9398	0.0036	0.9400	0.9438	1.0329	1.0311	1.0285	1.0401	0.9388	1.0308	1.0050
E2	1.0308	1.0355	1.0255	1.0315	0.9489	0.9484	0.9462	0.0037	0.9420	1.0446	1.0414	1.0411	1.0332	0.9469	1.0330	1.0110
E3	1.0320	1.0365	1.0261	1.0332	0.9508	0.9486	0.9477	0.9401	0.0038	1.0440	1.0405	1.0403	1.0417	0.9475	1.0308	1.0083
J1	1.0487	1.0567	1.0485	1.0535	1.0399	1.0415	1.0330	1.0407	1.0479	0.1022	0.9717	0.9687	0.9928	1.0039	1.0182	1.0176
J2	1.0476	1.0532	1.0473	1.0523	1.0330	1.0337	1.0309	1.0357	1.0387	0.9720	0.1002	0.9438	1.0029	1.0209	1.0245	1.0187
J3	1.0438	1.0562	1.0519	1.0549	1.0215	1.0262	1.0286	1.0299	1.0427	0.9675	0.9420	0.1026	0.9992	1.0028	1.0255	1.0241
J4	1.0434	1.0527	1.0457	1.0516	1.0195	1.0262	1.0304	1.0329	1.0413	0.9931	1.0035	0.9996	0.1070	1.0003	1.0276	1.0237
Sp	1.0235	1.0302	1.0221	1.0285	0.9483	0.9472	0.9476	0.9480	0.9512	1.0250	1.0023	1.0027	1.0010	0.0037	1.0318	1.0056
V1	1.0748	1.0762	1.0795	1.0786	1.0254	1.0266	1.0307	1.0296	1.0325	1.0182	1.0244	1.0253	1.0275	1.0203	0.1040	1.003
V2	1.0647	1.0560	1.0668	1.0591	0.9955	0.9950	0.9980	1.0017	1.0029	1.0185	1.0190	1.0246	1.0241	0.9940	1.0034	0.0495

Table A.8: **Pairwise normalized compression distance of annual movement data of 16 animals across 6 species.**

A.2 Synthetic Experiments Appendix A

A.2.1 Case Study: Shannon entropy as a tool for characterizing movement patterns of sheep with neurodegenerative disease

I created two synthetic agents here to represent the abnormal and normal sheep and 10 generalized pareto distributions with 10000 random values and with the parameters below:

$tail\ parameters = 0.1 : 0.1 : 1$

$scale\ parameter = 0.2$

$$shape\ parameters = \frac{0.2}{tail\ parameters}$$

The experiment was designed over 10 iterations where at each iteration and any point in time, the normal synthetic sheep can sample its distance covered from any of the ten distributions whereas the abnormal synthetic sheep can sample only from two of the distributions selected at every iteration at any point in time. The decision to limit the number of distributions for the abnormal sheep at each iteration was to simulate their tendency to get stuck in a behaviour compared to their normal counterparts. The sheep can only sample 1000 times per iteration per day. I computed the entropy over the 10 iterations and averaged the result.

A.2.2 Case study: Kolmogorov complexity as tool for classifying animal movement patterns across scales

I carried out two sets of experiments here. One with a small amount of data (720 location points) and another one with a larger amount of data (8760 points) to simulate the movement for one month and a year respectively. I generated 10000 data points from 10 bivariate gaussian distributions with means:

$$mean\ (\mu) = \begin{bmatrix} 0.001 & 0.001 \\ 0.01 & 0.01 \\ 0.1 & 0.1 \\ 1 & 1 \\ 10 & 10 \\ 15 & 15 \\ 20 & 20 \\ 25 & 25 \\ 30 & 30 \\ 35 & 35 \end{bmatrix}$$

and covariance matrix (CM):

$$CM = \begin{bmatrix} \mu + \mu/4 & \mu + \mu/4 \\ \mu + \mu/4 & \mu + \mu/4 \end{bmatrix}$$

I used the same covariance for all. 50 synthetic animals were created overall and divided into a group of 10, 5 per group. Each member of the group sampled its velocity from the same bivariate gaussian distribution. I computed the distance covered and also the pairwise NCD afterwards. The results, Figure A.3 and Figure A.4 show that the 10 groups of synthetic animals can easily be separated using this approach.

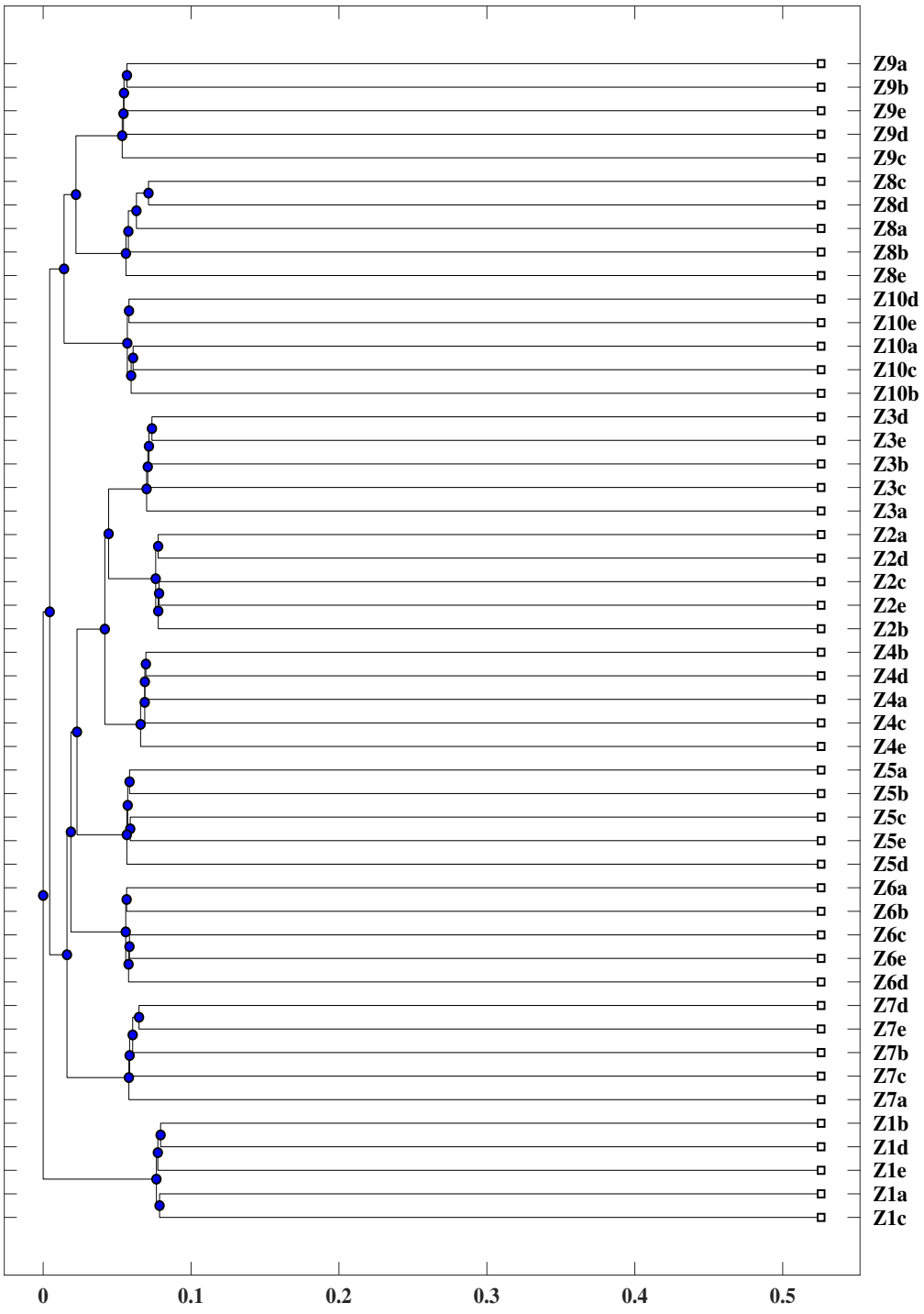


Figure A.3: **Hierarchical clustering of the pairwise NCD of the distance covered by 50 synthetic animals whose velocities were drawn from 10 distributions representing the movement patterns over a period of one year.** Note: Z1b for example indicates a synthetic animal is the second to sample its velocity from the first random distribution for velocity.

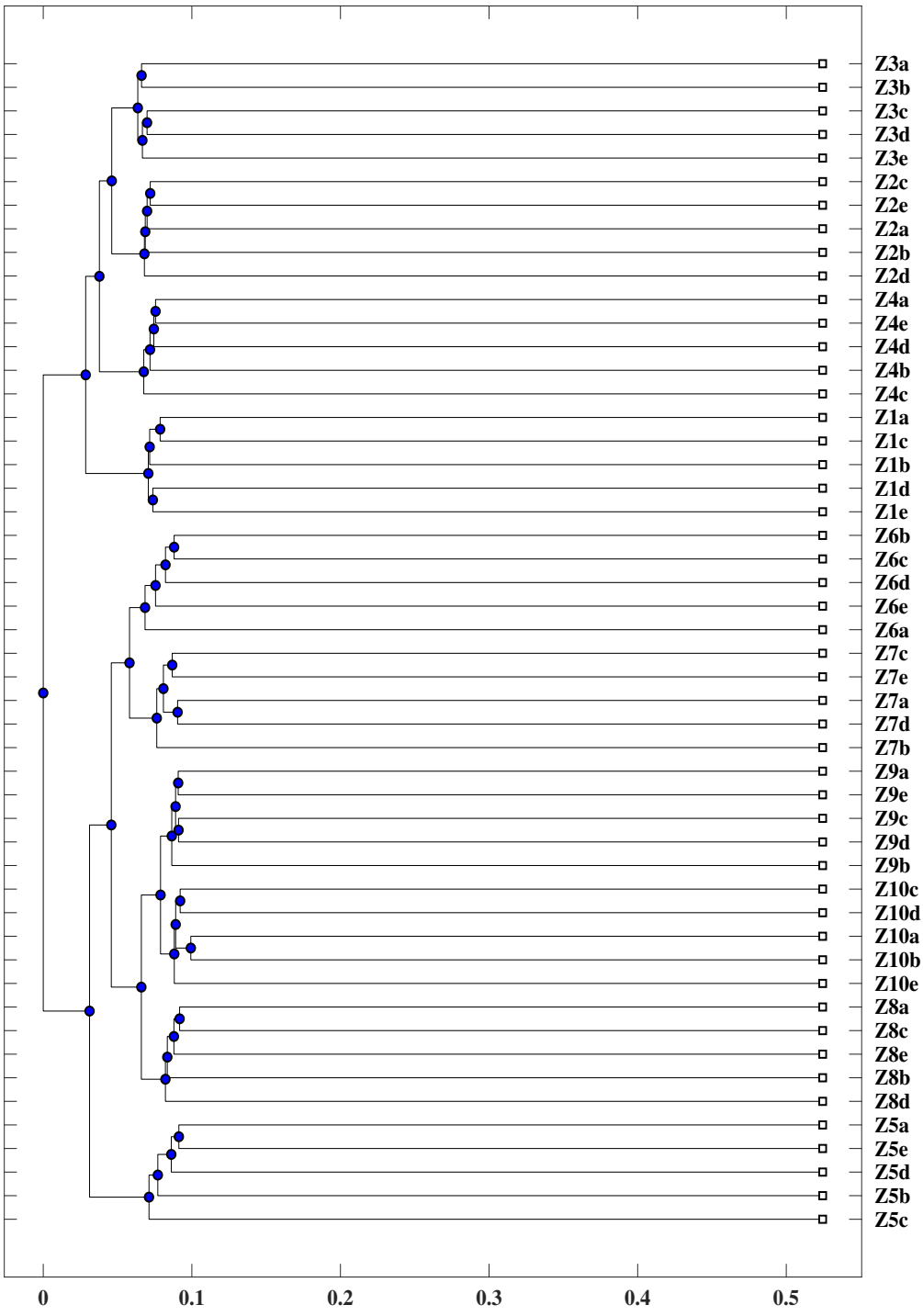


Figure A.4: **Hierarchical clustering of the pairwise NCD of the distance covered by 50 synthetic animals whose velocities were drawn from from 10 distributions representing the movement patterns over a period of one month.** Labels are as described above.

A.3 Ethics Statements Appendix A

A.3.1 Paper: A Significance Test for Inferring Affiliation Networks from Spatio-Temporal Data.

Ethics statement: Eleven (5 Batten disease ewes carrying a mutation in CLN5 and 6 healthy ewes age matched control) with mean age of 2 years were obtained from Lincoln University, New Zealand (NZ) where they have been reared under appropriate procedures approved by the Lincoln University Animal Ethics Committee in compliance with the NZ Animal Welfare Act (1999) and in accordance with US National Institutes of Health guidelines. The sheep had water and grazing available to them all through the period of the study.

A.3.2 Paper: Temporal and contextual consistency of leadership in homing pigeon flocks.

Ethics statement: The experiments described in this study were conducted according to Swiss regulations on animal welfare and experimentation, license 92/2011 issued by the Zurich Cantonal Veterinary Office. This license was approved by an ethics committee (composed by scientists and animal protection organizations), which is responsible for animal welfare and experimentation licensing for all universities in the canton of Zurich. The license generically approves GPS tracking of homing pigeons, which is considered non-invasive and not imposing stress on the animals (severity level 0 according to Swiss classification). The owner of the pigeons is a co-author in this study (H.-P.L.), the lofts being located on his property. The lofts were approved by the Cantonal Veterinary Office as facility for experimental animals.

A.3.3 Paper: Environmental drivers of variability in the movement ecology of turkey vultures (*Cathartes aura*) in North and South America.

Note: Ethics statement not available.

A.3.4 Paper: Vegetation dynamics drive segregation by body size in Galapagos Tortoises migrating across altitudinal gradients.

Ethics statement: All animal handling procedures for this study adhered to the policies of the Galapagos National Park and the Max Planck Institute of Ornithology.

A.3.5 Paper: Home range plus: a space-time characterization of movement over real landscapes.

Ethics statement: This research was approved by the UC Berkeley Animal Care and Use Committee R217-0509BR217-0509B.

A.3.6 Paper: Methods for assessing movement path recursion with application to African Buffalo in South Africa.

Note: Ethics statement not available.

A.3.7 Paper: Black- backed jackal exposure to rabies virus, canine distemper virus and bacillus anthracis in Etosha National Park, Namibia.

Ethics statement: Twenty-two jackals were fitted with GPS VHF collars recording GPS locations hourly for 1-2 yr. All animals were captured and released safely. This animal handling protocol was approved under University of California, Berkeley ACUC R217-0509B and by the Namibian Ministry of Environment and Tourism Game Capture Veterinarian.

A.3.8 Paper: The environmental-data automated track annotation (env-data) system: Linking animal tracks with environmental data.

Ethics statement: Galapagos Albatross data used for the case study were collected with permission and support of the Galapagos National Park Service.

A.3.9 Paper: Where do livestock guardian dogs go? movement patterns of free-ranging maremma sheepdogs.

Ethics statement: All research was carried out in compliance with the Australian Code for the Care and Use of Animals for Scientific Purposes, 7th edition. Ethics approval was obtained from the Animal Ethics Committee of the University of Tasmania (approval number: A0012323).

A.3.10 Paper: Foraging Behavior and Success of a Mesopelagic Predator in the Northeast Pacific Ocean: Insights from a Data-Rich Species, the Northern Elephant Seal.

Ethics Statement : The animal use protocol for this research was reviewed and approved by the University of California at Santa Cruz Institutional Animal Care and Use Committee and followed the guidelines established by the Canadian Council on Animal Care and the ethics committee of the Society of Marine Mammalogy. Research was carried out under National Marine Fisheries Service permits: # 786-1463 and #87-143.

A.3.11 Paper: Factors influencing foraging search efficiency: Why do scarce lappet-faced vultures outperform ubiquitous white-backed vultures.

Ethics Statement : All experiments carried out in accordance to the South African Bird Ringing Unit protocol.

A.3.12 Paper: In search of greener pastures: Using satellite images to predict the effects of environmental change on Zebra migration?

Ethics Statement : All animal handling procedures were covered by a University of Bristol investigation number (UB/06/012) and by a Botswana research permit (EWT3/3/8XXXIV).

Appendix B

Quantifying Abnormal Movement Behaviour in sheep with Batten disease (Supplementary Material)

B.1 Supplementary Results

		Sheep ID										
Sheep ID	Batten Disease	1	2	3	4	5	6	7	8	9	10	11
1	No	-	NS	NS	NS	NS	NS	NS	NS	NS	S	NS
2	No	NS	-	NS	NS	NS	S	NS	NS	S	S	NS
3	No	NS	NS	-	S	NS	NS	NS	NS	NS	S	NS
4	No	NS	NS	S	-	NS	NS	NS	NS	S	S	NS
5	No	NS	NS	NS	NS	-	NS	NS	S	S	S	NS
6	No	NS	S	NS	NS	NS	-	NS	S	S	S	S

Table B.1: **Pairwise significance when K-means was used.** Where S = Significant when p-value $\leq$ 0.05, and NS = Not significant when p-value $>$ 0.05.

		Sheep ID										
Sheep ID	Batten Disease	1	2	3	4	5	6	7	8	9	10	11
1	No	-	NS	NS	NS	NS	NS	NS	S	S	S	S
2	No	NS	-	NS	NS	NS	NS	S	S	S	S	NS
3	No	NS	NS	-	NS	NS	NS	NS	NS	S	S	NS
4	No	NS	NS	NS	-	NS	NS	NS	S	S	S	NS
5	No	NS	NS	NS	NS	-	NS	NS	S	NS	S	NS
6	No	NS	NS	NS	NS	NS	-	S	S	S	S	S

Table B.2: **Pairwise significance when gaussian based model was used.** Where S = Significant when p-value ≤ 0.05, and NS = Not significant when p-value > 0.05.

		Sheep ID										
Sheep ID	Batten Disease	1	2	3	4	5	6	7	8	9	10	11
1	No	-	NS	NS	NS	NS	NS	S	S	NS	NS	NS
2	No	NS	-	NS	NS	NS	NS	NS	NS	NS	NS	S
3	No	NS	NS	-	NS	NS	NS	NS	NS	NS	NS	S
4	No	NS	NS	NS	-	NS	NS	NS	S	S	NS	S
5	No	NS	NS	NS	NS	-	NS	NS	NS	NS	NS	NS
6	No	NS	NS	NS	NS	NS	-	NS	NS	NS	NS	NS

Table B.3: **Pairwise significance when gaussian based model with log of distance was used.** Where S = Significant when p-value ≤ 0.05, and NS = Not significant when p-value > 0.05.

		Sheep ID										
Sheep ID	Batten Disease	1	2	3	4	5	6	7	8	9	10	11
1	No	-	NS	NS	NS	NS	NS	NS	NS	S	S	NS
2	No	NS	-	NS	NS	NS	NS	NS	S	NS	S	NS
3	No	NS	NS	-	NS	NS	NS	NS	NS	S	NS	NS
4	No	NS	NS	NS	-	NS	NS	NS	NS	NS	S	NS
5	No	NS	NS	NS	NS	-	NS	NS	S	NS	S	NS
6	No	NS	NS	NS	NS	NS	-	NS	NS	S	S	NS

Table B.4: **Approach2: Pairwise significance with GMM.** Where S = Significant when p-value ≤ 0.05, and NS = Not significant when p-value > 0.05.

		Sheep ID										
Sheep ID	Batten Disease	1	2	3	4	5	6	7	8	9	10	11
1	No	-	NS	NS	NS	NS	NS	NS	NS	S	NS	NS
2	No	NS	-	NS	NS	NS	NS	NS	NS	NS	S	NS
3	No	NS	NS	-	NS	NS	NS	NS	NS	NS	NS	NS
4	No	NS	NS	NS	-	NS	NS	NS	NS	NS	NS	NS
5	No	NS	NS	NS	NS	-	NS	NS	S	S	NS	NS
6	No	NS	NS	NS	NS	NS	-	NS	NS	S	S	NS

Table B.5: **Approach2: Pairwise significance with K-means.** Where S = Significant when p-value $\leq$ 0.05, and NS = Not significant when p-value > 0.05.

Appendix C

Learning to Herd (Supplementary Material)

C.1 More Results from Ablation Studies

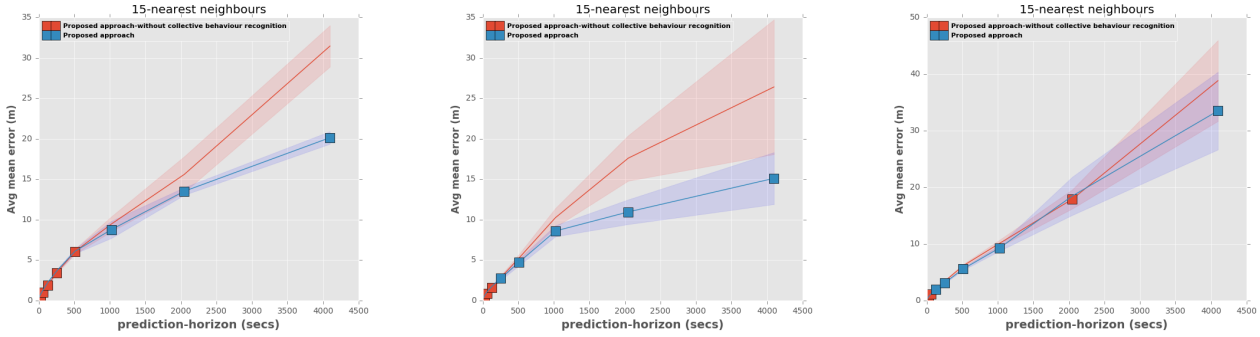


Figure C.1: **Ablation - Model -1 (No collective behaviour recognition) - Metric-1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. It can be seen that the collective behaviour recognition is important for downstream tasks. The markers are temporal intervals where the difference is significant using a Mann Whitney U test.

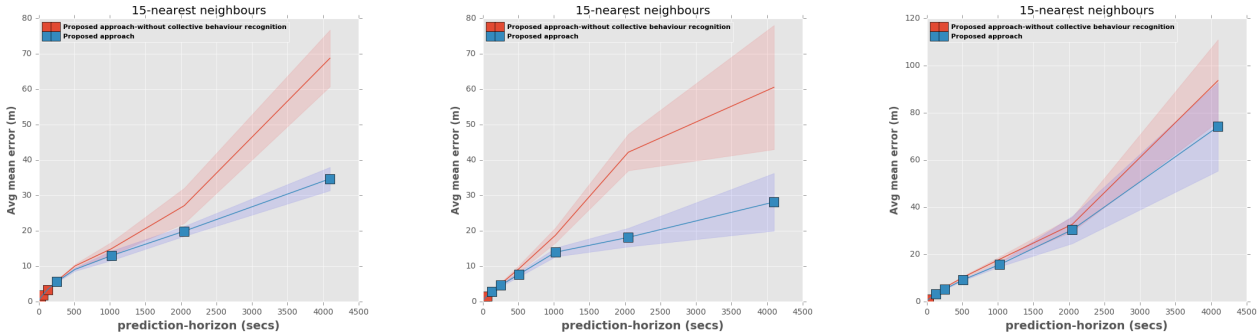


Figure C.2: **Ablation - Model -1 (No collective behaviour recognition) - Metric-2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. It can be seen that the collective behaviour recognition is important for downstream tasks. The markers are temporal intervals where the difference is significant using a Mann Whitney U test.

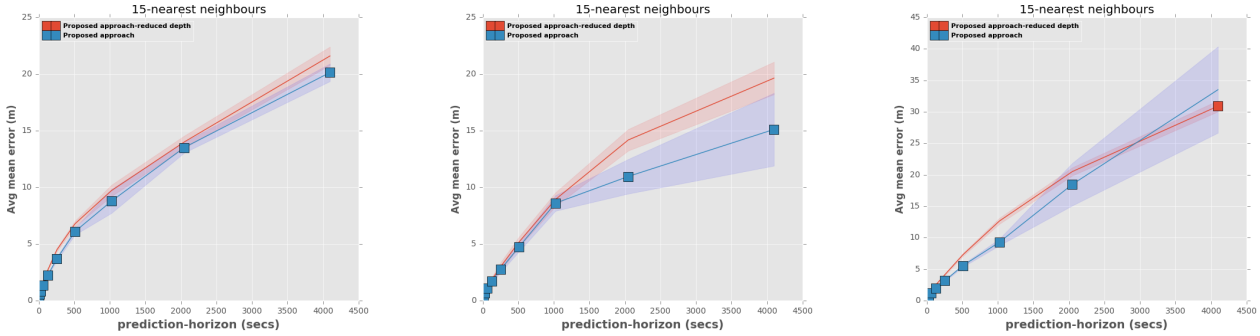


Figure C.3: **Ablation - Model - 2 (Reduced depth of the hierarchy) - Metric-1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. It can be seen that deep networks produced better results than a shallow one. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test.

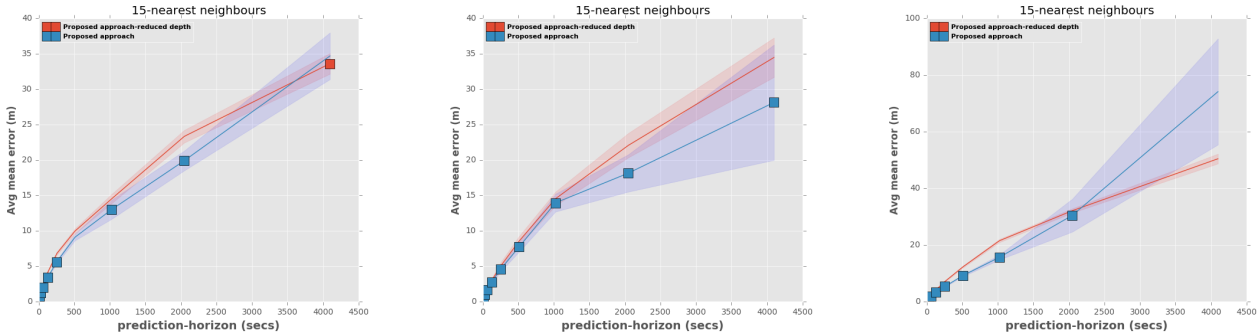


Figure C.4: **Ablation - Model - 2 (Reduced depth of the hierarchy) - Metric-2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. It can be seen that deep networks produced better results than a shallow one. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test.

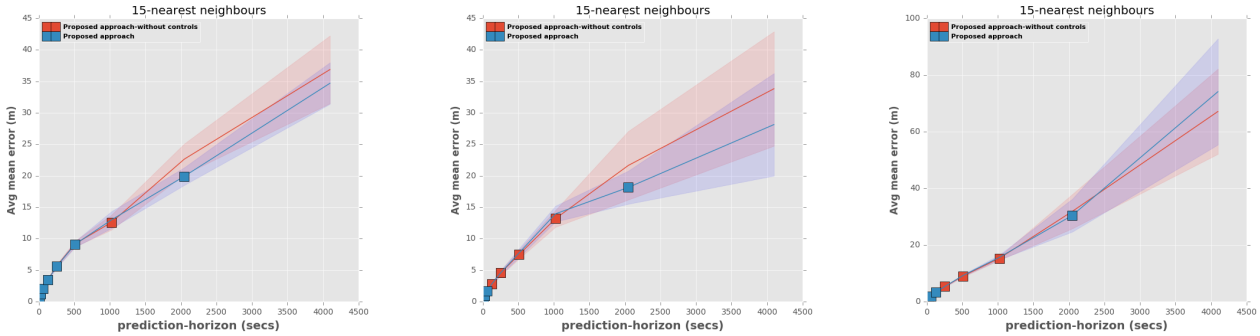


Figure C.5: **Ablation - Model - 3 (Without top layer control) - Metric-1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. The proposed approach can be seen to perform better than this model sometimes and some other times the ablation model performs better. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test.

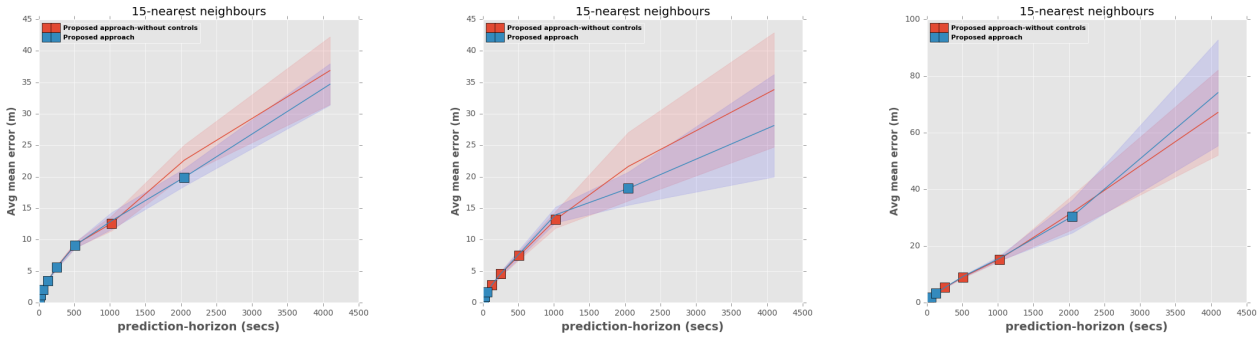


Figure C.6: **Ablation - Model - 3 (Without top layer control) - Metric-2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. The proposed approach can be seen to perform better than this model sometimes and some other times the ablation model performs better. The marker points are temporal intervals where the difference is significant using a Mann Whitney U test.

C.2 More Performance Comparison Results

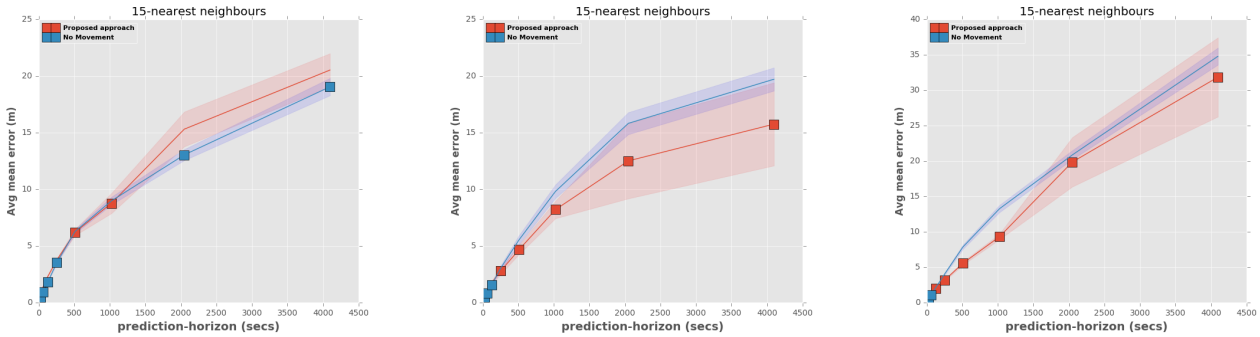


Figure C.7: **Performance comparison with respect to baseline 1 - no movement using metric 1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. Proposed approach can be seen to outperform no movement baseline most of the time. The markers are areas where the difference is significant using a Mann Whitney U test.

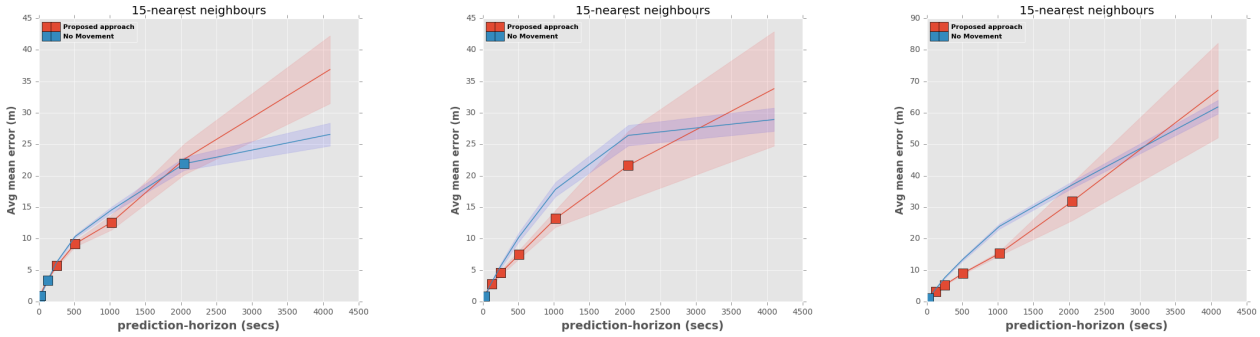


Figure C.8: **Performance comparison with respect to baseline 1 - no movement using metric 2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. Proposed approach can be seen to outperform no movement baseline most of the time. The markers are areas where the difference is significant using a Mann Whitney U test.

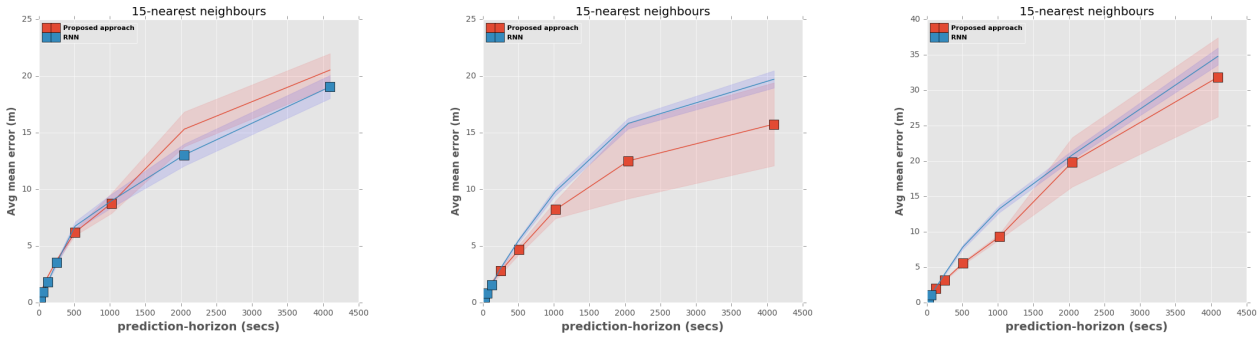


Figure C.9: **Performance comparison with respect to baseline 2 - RNN using metric 1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. Proposed approach can be seen to outperform RNN most of the time. The markers are areas where the difference is significant using a Mann Whitney U test.

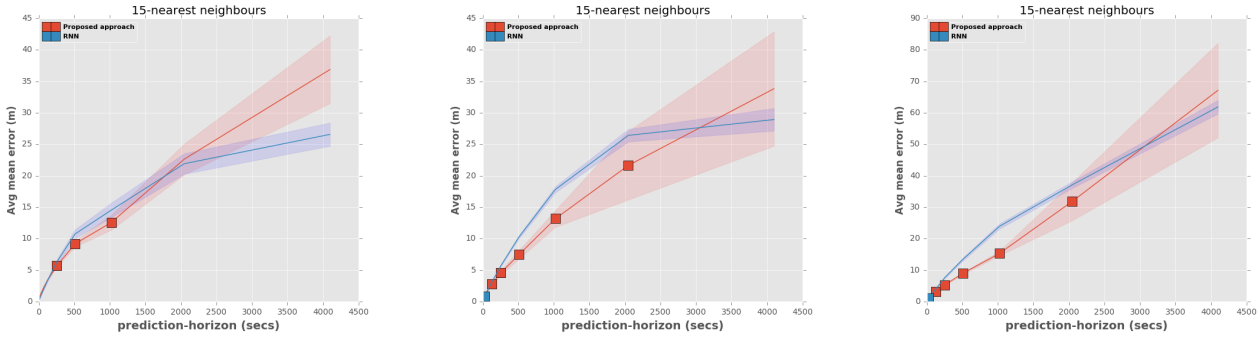


Figure C.10: **Performance comparison with respect to baseline 2 - RNN using metric 2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. Proposed approach can be seen to outperform RNN most of the time. The markers are areas where the difference is significant using a Mann Whitney U test.

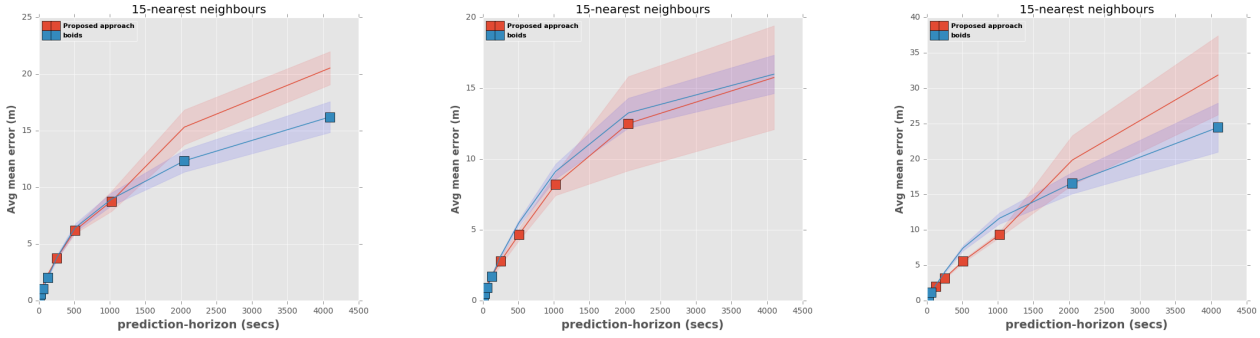


Figure C.11: **Performance comparison with respect to baseline 3 - boids using metric 1.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. The performance of each approach varies depending on the dataset with no clear separation on the winner. The markers are areas where the difference is significant using a Mann Whitney U test.

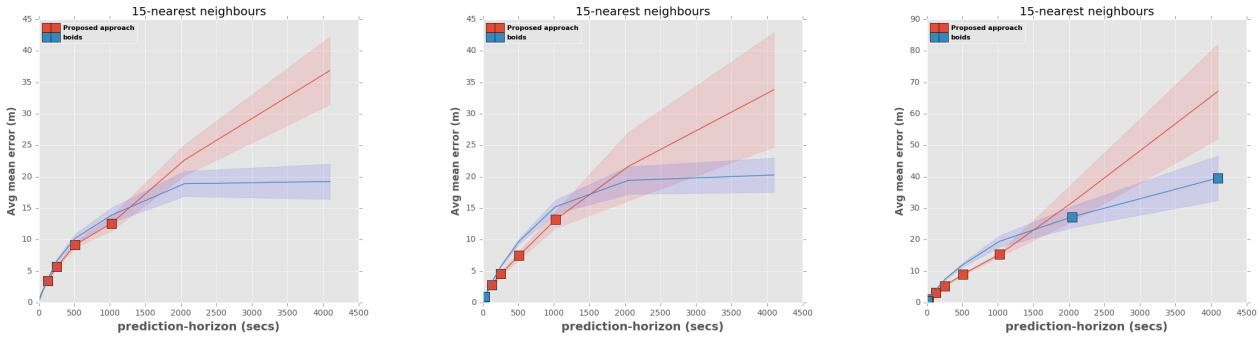


Figure C.12: **Performance comparison with respect to baseline 3 - boids using metric 2.** Left: Test dataset 1, Centre: Test dataset 2, Right: Test dataset 3. Boids can be seen to outperform the proposed approach over long horizons and vice-versa over short horizons. The markers are areas where the difference is significant using a Mann Whitney U test.

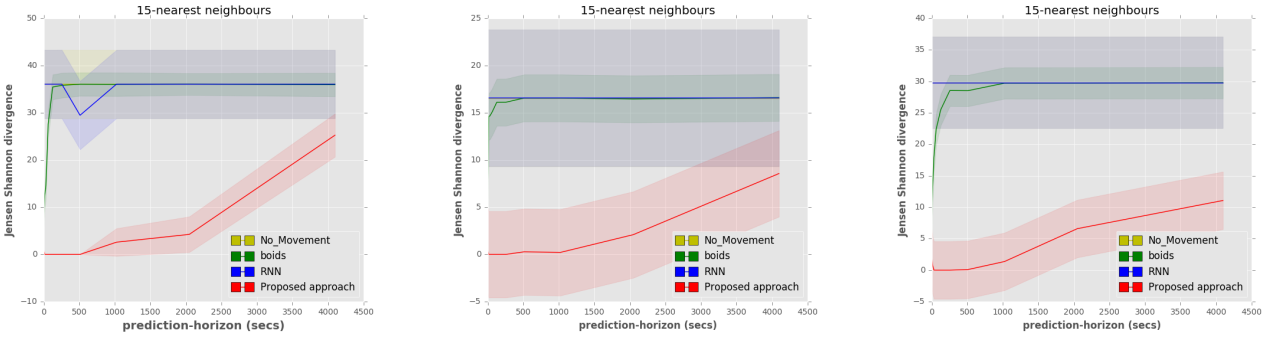


Figure C.13: **Performance comparison with respect to all baselines using metric - 3.** It can be seen that the proposed approach outperform all baselines.