Radboud University

DONDERS INSTITUTE

RADBOUD UNIVERSITY

MASTER THESIS

Computerized Behavioural Analysis in Mice

At
Memory Dynamics
in
Neuroinformatics

July 8, 2019

Author:
Steven Smits

Supervisor:
Dr. F.P. (Francesco) Battaglia

Abstract

Autism is a cluster of behavioural abnormalities that manifest as impaired social behaviour, perseverance behaviours and altered memory processes. The study of memory processes in mice is used as a model for normal and pathological cognitive functioning. The euchromatic methyltransferase 1 heterozygous knockout (ehmt1$^{+/-}$) mouse is a model for Kleefstra syndrome, a condition characterized by autism and intellectual disability. Studies on memory processes in ehmt1$^{+/-}$ have mainly focused on episodic memory, with mixed results on whether this is improved or equal to healthy subjects. In this study, ehmt1$^{+/-}$ mice were subjected to the Object-Space task, a novel paradigm to distinguish episodic from semantic-like memory. In this task mice are exposed to multiple trials involving objects placed dynamically but with overlapping regularity (overlapping condition), objects placed all in the same location over trials (stable), or objects randomly placed each trial. Over trials, mice may acquire the spatial patterns in the first two conditions but not last. However, one major challenge is extracting comprehensive behavioural information from video data of mice performing such a task. This thesis describes a computerized method to categorize various behaviours (i.e. Object Exploration, Wall Exploration, and Corner Sitting) from video data of mice performing the Object-Space task. The method involves a model that uses techniques such as kinetic action recognition, transfer learning, and pose estimation to categorize behaviours in both a supervised and an unsupervised manner. The former implements optic flow over multiple frames in order to learn what constitutes a behavioural module of Object Exploration. The latter implements recent developments in deep learning for pose estimation to define both Wall Exploration and Corner Sitting behaviours as a geometrical configuration of limbs. Visual inspection of these models combined show it to be highly accurate in time in terms of sensitivity and specificity of action classification. Moreover, this behavioural categorization model was used to describe an array of behaviours (e.g. object exploration time, object discrimination index) of mice performing trials in all conditions of the Object-Space task. This array of behaviours could be used to predict genotype (i.e. ehmt1$^{+/-}$ or ehmt1$^{+/+}$) of a mice based on a single video of a trial in both the overlapping and stable condition, but not in the control condition. One especially interesting finding is that the models to predict genotype used more memory related behaviours (e.g. discrimination index) to predict genotype in the overlapping condition, whereas the models to predict genotype in the stable condition mainly used general behaviours (e.g. total exploration time). Further inspection of these behaviours between genotypes show that ehmt1$^{+/-}$ mice may display increased memory expression behaviours over healthy controls. This indicates that memory processes in this Kleefstra mice model might be improved, and not characterized by intellectual disability as previously thought.

Contents

Introduction

Innate behaviours allow animals to attain goals such as obtaining food or defending from a predator. The study of ethology proposes that simple action sequences form modules of coherent behaviour, which is all embedded into neural circuits (Tinbergen, 1951). This idea has led translational neuroscience researchers to use animal behaviour as a proxy of what they're really interested in, i.e. neurological mechanisms that underlie these behaviours (Baker, 2011). For example, Yizhar et al. (2011) showed that optogenetically causing an excitation/inhibition balance in the rodent medial prefrontal cortex concomitantly causes rodents to reduce social behaviour in a social exploration task. As neurological manipulation becomes increasingly available and research accumulates, there is need for behavioural methods that surrogates neurological processes. Moreover, behavioural analysis of such a method needs to be reliable, invariant of the assessor. Current standards of categorizing rodent behaviour are based on test batteries that require human visual inspection of the video data (Blanchard, Griebel, and Blanchard, 2003; Pandey et al., 2008). Assessing behaviour by means of visual inspection may introduce complications due to a possible large amount of behavioural categories to assess, sequential complex actions that could arise from simpler actions, and ambiguous situations. All of these factors reduce inter/intra-rater reliability. Prior work has addressed these issues by using automated techniques that are able to track and categorize rodent behaviour based on video data (De Chaumont et al., 2012; Wiltschko et al., 2015). These models could, for example, take into account specific body parts (e.g. tail, body, head) of the rodent in order to track its position, orientation and speed. These parameters could then be used to define relative position to other objects or rodents to assess if the tracked rodent is interacting with it. Dere, Huston, and Silva (2005), for example, used geometrical primitives (e.g. circles) to model mice and track mice. They then used this model to describe a repertoire of behaviours. Their method used constraints and physics engines to solve for adaptive body parts movement, which sometimes produces errors of body parts switching in assigned location. In another study by Wiltschko et al. (2015), behavioural repertoires of mice were extracted as stereotypies by using an autoregressive hidden markov model. Their method was able to find previously unexplored latent behaviours of mice, although may not always be useful when classifying a priori defined behaviours. Together, these studies show that automatically categorizing behaviours of mice is plausible by applying machine learning techniques. Notwithstanding the current state of the art, there is no model that is reliably able to track multiple body parts (e.g. ears, nose, tail) of a rodent and extracts behavioural information.

This thesis describes a novel method based on a multitude of advances in machine learning that extracts behavioural states in mice using temporal and pose information. The developed method includes tracking multiple limbs, inferring head direction, and uses the temporal evolution of the mouse's movement to derive three main behavioural categories: object exploration, wall exploration and corner sitting. Furthermore, this method has a frame-wise resolution which allows it to describe these behaviours and its relations, such as transition probabilities, as an unfolding evolution over time. To validate the usability of the developed method, it was used to predict whether a mouse was exploring an object over the course of time. This information was then used to describe multiple behavioural features (e.g.transition probabilities) of mice with either an autism-related gene dysfunction or healthy mice performing a memory task to be described. These behavioural features were then successfully used to predict the genotype of mice, showing memory related behaviours to be of importance for its feasibility.

1.1 Background

1.1.1 Automatic Behavioural Scoring

This study aims to develop an automated rodent behavioural classifier. The proposed approach will combine techniques of kinetic action recognition, transfer learning, and pose estimation. Combining these concepts provides a novel approach to potentially supersede previous attempts at computerized video analysis.

To the end of finding a function that complies with the mapping of video information to the behavioural dimension, one may be looking to existing classification approaches. Artificial neural networks (ANNs) have been shown to be useful as a method for classification problems in various fields (Bala and Kumar, 2017; Schmidhuber, 2015). This computational model is inspired by the biological neural networks that comprise the human brain in that a multitude of interconnected neurons are modifying their connections in response to some input that is experienced (van Gerven, 2017). The basic unit of an ANN is the neuron, which receives the input and transmits output as a mathematical function of the input:

$$y = f(\sum_{i=1}^{N} w_i x_i)$$

Where x_i is the input neuron, w are the synaptic weights, y is the output and N the number of neurons. This model is often extended to contain multiple hidden layers of neurons a_j, before generating a final output y. In the classification problem, the output y will ideally be equal to target value t that is the label of the class to which x belongs. In reality, the true function that transforms the input to a target output is unknown. By updating the synaptic weights between neurons, the ANN multi-layer perceptron procedure can approximate any continuous function (Hornik, 1991; Scarselli and Tsoi, 1998). The main approach to find the optimal synaptic weights to estimate an unknown function is to minimize some cost function (Nielsen, 2015). One example cost function is the binary Cross-Entropy loss function:

$$\mathcal{L}_t(y) = -(t \cdot log(y) + (1-t) \cdot log(1-y))$$

Where the cost $\mathcal{L}_t$ is zero if the predicted label y matches the target label t and the cost increases logarithmically by any increase in predictive deviation. With an unknown true transformation function and increasing number of synaptic connections, the minima of this cost function cannot be found analytically. For this, techniques such as gradient descent are often used. Altogether, the network will learn to use underlying features belonging to classes in order to predict the correct class y based on input x.

Kinetic Action Recognition With the vast amount of very deep ANNs (LeCun, Bengio, and Hinton, 2015) trained on big image classification data, one benefit is the use of existing architectures for other domains, such as action recognition in video data (Carreira and Zisserman, 2017). Carreira and Zisserman (2017) used the realization that deep networks can be trivially inflated to become spatio-temporal feature extractors with initialized weights of the former deep network. Here, inflating a network means modifying the architecture such that the input allows for a sequence of images (i.e. video) as opposed to a single image. The authors showed that inflating an Inception network (Szegedy et al., 2015) and subsequently training it on the Kinetics Human Action Video Dataset (Kay et al., 2017) achieves top performance in both action recognition and image classification. Inception networks include so called Inception modules, where each module consists of parallelized different sized convolutional filters followed by batch normalization and some activation function (e.g. ReLu or softmax) stacked and concatenated as one high-dimensional output for the next layer. Figure 1.1 depicts how this was implemented by Carreira and Zisserman (2017). An Inception network architecture will then be composed by varying convolutional layers, max pooling layers, Inception modules and activation functions.

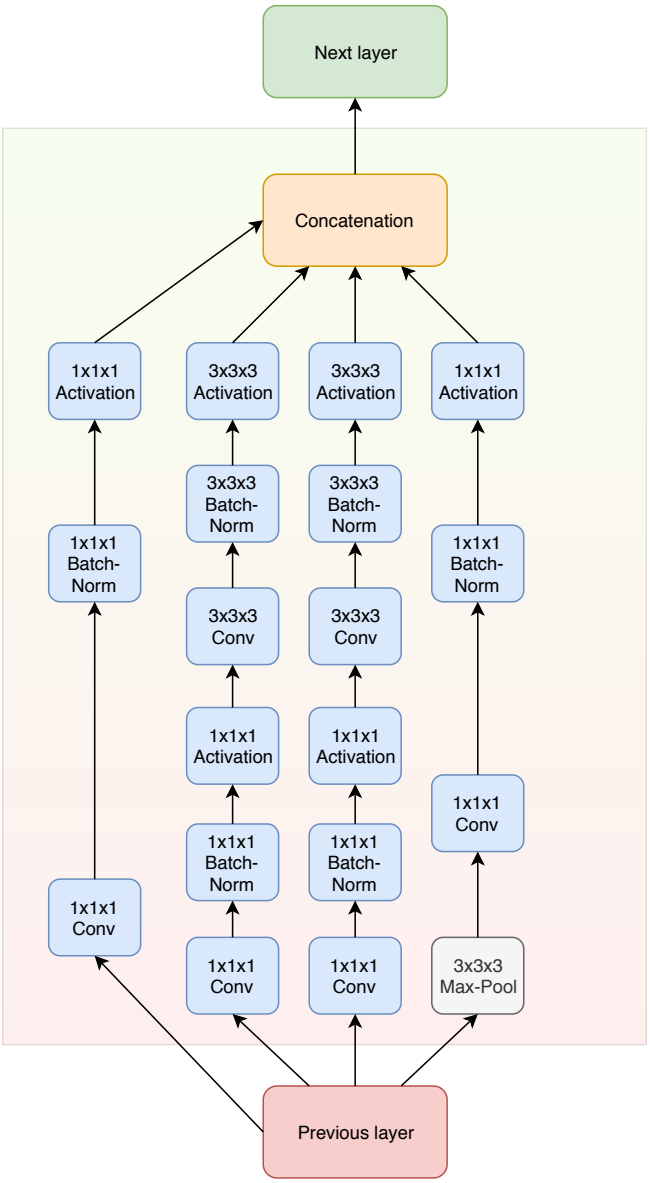


FIGURE 1.1: Inception module (Inc.) of an inflated network suited for videodata. The module includes parallel stacked different sized convolutional (conv) filters, ending with an activation layer (e.g. ReLU, softmax)

Transfer Learning One phenomenon in deep ANNs trained on images is that lower layers show general features (e.g. color blobs and Gabor filters) across all kind of datasets, and higher layers show task-specific features (e.g. mouse). This fact is widely used in the field to apply a technique called transfer learning, which is the idea that one trained ANN can be used as a starting point to train another (Yosinski et al., 2014). Specifically this means that the first n layers and their weights of some trained network are copied as initializers for another target network, where this target network is subsequently trained on the desired classification task. Transfer learning has been shown to speed up training and improve performance of the deep ANN for the desired target task in many domains (Mittal, Vatsa, and Singh, 2015; You et al., 2015; Long et al., 2015).

Pose Estimation Quantifying behaviour may be achieved by extracting the geometrical configuration of multiple body parts (i.e. pose), and its relation to specified actions. Mathis et al. (2018) applied the concept of transfer learning on deep ANNs to extract the pose of many types of animals in single

images (Nath* et al., 2018). Specifically, Mathis et al. (2018) made an ANN that has feature layers from the DeeperCut (Insafutdinov et al., 2016) and outer layers trained to estimate the pose of animals. The authors' have shown that their method produces state-of-the-art computerized pose estimation without the need to physically mark the animals.

1.1.2 Mouse Autism Model

This study will aim to apply the developed automated behavioural classifier method to investigate memory processes in mice. It will specifically investigate episodic and semantic processes in a mouse model of autism.

Memory Mechanisms The ability to acquire knowledge about the world is of essence for survival in the animal kingdom. For example, during a period of drought an elephant may recall the location of a water source in a similar context when it was young. Armed with the power of this memory, it could lead its herd to successful hydration, and effectively increasing their likelihood of survival. But what if this time the previous abundant source is now devoid of water? Shall the elephant survive and live to find itself in a similar situation once more, should it recall that the source had water for many years or should it recall that the source if now empty? This question marks an important distinction of aspects in the mechanism of memory storage. In semantic memory, general statistical attributes are extracted cumulatively across a multitude of events (Squire, 2004). Conversely, in episodic memory features of specific events are retained. As memory consolidates, one stored event may transition between these two types of memory mechanisms (Frankland and Bontempi, 2005; Moscovitch et al., 2016). Differentiation between these two mechanisms of memory has especially proven to be challenging in rodent studies, disproportionately focusing more on episodic rather than semantic memories (Dere, Huston, and Silva, 2005; Roberts, 2016).

Kleefstra Model Autism spectrum disorder is a cluster of behavioural syndromes characterized by early childhood onset of neurodevelopmental abnormalities including impairments in social interactions and communication, and a restricted range of interests, often associated with repetitive and stereotyped behaviors (Klinger et al., 2019). Notably, memory processes are often affected in individuals with autism spectrum disorder. Specifically, individuals with autism spectrum disorder may have an impaired episodic memory, yet an unaltered or enhanced semantic memory (Goddard et al., 2014; Gaigg, Bowler, and Gardiner, 2014). Studying autism independently is hard as individuals with this disorder tend to have a conglomeration of disorders and symptoms. Kleefstra syndrome is a genetic disorder characterized by features of autism, and other abnormalities (Kleefstra et al., 2006). This syndrome is mainly caused by haploinsufficiency of euchromatic histone methyltransferase 1 (ehmt1), a regulator of synaptic scaling that is critical for neurodevelopmental aspects such as neural network activity (Benevento et al., 2016; Balemans et al., 2012). Studies using mice with a heterozygous ehmt1 (ehmt1$^{+/-}$) gene have shown them to have reduced exploration and increased anxiety in novel environments, and increased pattern separation in other contexts (Balemans et al., 2010; Benevento et al., 2017). With the scarce amount of studies on semantic memory abilities in rodents (with and without autism), varying memory mechanisms may need refined characterization.

Methods

2.1 Dataset

The dataset used throughout this thesis consists of videos of mice performing the Object-Space task as described below. Most of these videos are scored by humans on mice either exploring an object or not.

2.1.1 Subjects

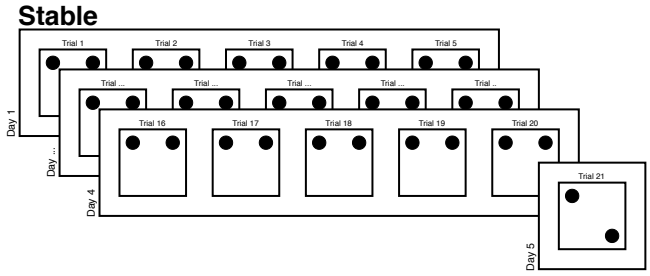
Male wildtypes ehmt1$^{+/-}$ and ehmt1$^{+/+}$ and mice (bred in-house), 12-16 weeks of age at the start of behavioural training were group housed with ad libitum access to food and water. Mice were maintained on a 12 hour dark-light cycle and tested during the light period. In compliance with Dutch law and Institutional regulations, all animal procedures were approved by the Central Commissie Dierproeven (CCD) and conducted in accordance with the Experiments on Animal Act.

2.1.2 Object-Space Task

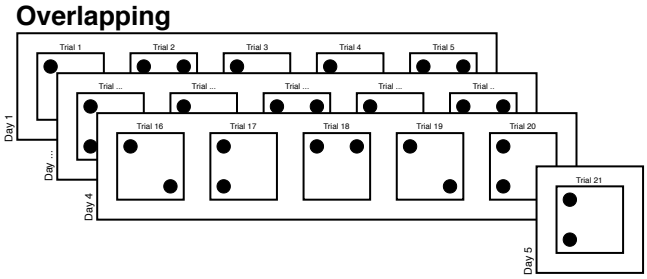
The Object-Space task, developed by Genzel et al. (2018), aims to distinct between episodic and semantic memory using behaviour as a proxy. For the full habituation and training procedure, please refer to Genzel et al. (2018).

In the Object-Space task, mice are allowed to explore two objects in a 75cm$\times$75m box with either a white or a green background (see Figure 2.1a. Objects could either be in the upper right, lower right, lower left, or upper left. The general reasoning is that mice tend to prefer exploring objects at novel locations. For a total of 21 trials, objects are placed with different patterns that could be either *stable*, *overlapping*, or *randomly* configured. For all conditions mice are trained for 5 trials, each 5 minutes, per day for 4 days with object types always being identical within a trial but varying between trials. Subsequently, 24 hours after the last training trial (trial 20), a test trial with a condition-specific configuration is done for 10 minutes. In the stable condition (see Figure 2.1b), objects are always at the same location for the first 20 trials. Then, at the test trial one object is moved to a novel location. With one object moved at the test trial, it is expected that mice explore the object at the novel location more than the static one. Moreover, "knowing" which location is the novel could either be solved by remembering the last training trial, or forming some cumulative memory of all training trials. In the overlapping condition (see Figure 2.1c), one object is always presented at the same location across the first 20 trials whereas the other varies at the other three potential locations. Then, the configuration of objects is the same in the test trial as in trial 20. This is paramount for distinguishing between memory processes. To elucidate, if only the last training trial is remembered then there is no novel object location and no object preference is expected in exploration. In contrast, if the mouse has acquired the cumulative statistical knowledge from the training trials that one location always has an object placed on it then the other object in will be relatively novel in trial 21. Thus, with some semantic process the mouse may explore the object at the varied location more than the object at the stable location. In the random condition (see Figure 2.1d), object placement is psuedorandomly configured across training and test trials such that there ought to be no perceived pattern. With no pattern, no object location discrimination in exploratory behaviour is expected.

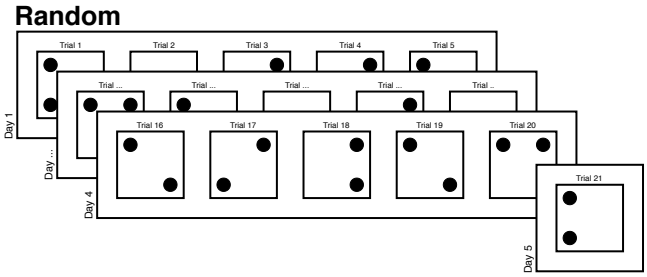
(A) An example box layout with objects in the upper right and lower left.



(B) Schematic example of a stable trial sequence. Here, for the first twenty trials all objects are in the same location. One object is moved in the last trial.



(C) Schematic example of an overlapping trial sequence. Here, for all trials one object remains at the same location. The other object varies in location for the first 19 trials, but is stable in the last two trials.



(D) Schematic example of a random trial sequence. Here, for all trials, all objects are pseudorandomly placed such that there is no spatial pattern.

FIGURE 2.1: Object-Space task for mice. One full sequence of 21 trials is a week's session that consists of 5 training trials per day for 4 days, and a probe trial 24 hours after the last trial on day 4.

2.2 Automatic Behavioural Scoring

The methods used to arrive at a fully automated behavioural scoring of video data with mice in the Object-Space task will be described in this section.

The data extracted from the Object-Space task consists of varying length videos $\mathcal{V}$ filmed from above a white or green square box that contains two static objects and one moving black mouse. The goal is to find some mapping per $\mathcal{V}$ such that the output is a time-series vector $\boldsymbol{x} = \begin{pmatrix} x_1 & \dots & x_t \end{pmatrix}$ of length t that per frame i outputs what action the mouse is doing x_i, with action set:

$$\chi = \{\text{exploring object}, \text{exploring wall}, \text{sitting in corner}\}$$

This is a function $f :: \left\{ \mathcal{V} \mapsto \boldsymbol{x} = \begin{pmatrix} x_1 & \dots & x_t \end{pmatrix} \right.$.

The implementation

To predict the actions of Objection Exploration, Wall Exploration, and Corner Sitting, two models will be presented. The first model will use Kinetic Action Recognition to extrapolate the Object Exploration behaviour of the mice. The second model will use pose estimation to extract the actions Wall Exploration and Corner Sitting. The reason for this split is due to in-lab availability of action labeled Object Exploration video data, whereas the other actions have no such dataset. In the end, the predicted actions of both models can be frame-wise concatenated for each video trial.

2.2.1 Object Exploration

The goal is to find some mapping per $\mathcal{V}$ such that the output is a time-series vector $\boldsymbol{x} = \begin{pmatrix} x_1 & \dots & x_t \end{pmatrix}$ of length t that per frame i outputs whether the mouse was exploring an object or not x_i, that is a function $f_E :: \left\{ \mathcal{V} \mapsto \boldsymbol{x} = \begin{pmatrix} x_1 & \dots & x_t \end{pmatrix} \right.$.

Dataset A dataset was created that consists of 100 stacked videos (width: 384, height: 512) of mice performing the Object-Space task in a white background box that is frame-wise manually labeled by humans for object exploration. In total this dataset consists of 910237 pairs of frames and target labels. Next, the dataset was pseudo-randomly split into a training (90%) and validation (10%) set. That is, batches of 27 frames were put into either the training or validation set.

The model A deep inflated inception ANN was made to classify videodata of mice performing the Object-Space task by applying transfer learning to the human action recognition model by Carreira and Zisserman (2017). The original model (see Figure 2.2a) was designed and trained to classify 400 human actions. Applying the ideas of transfer learning, the higher layers of this model were removed and replaced by several convolutional layers and a final activation layer to output 2 classes: exploring object and ¬exploring object (see Figure 2.2b). This means that the orange part in Figure 2.2b has initialized weights from the model trained by Carreira and Zisserman (2017) and the green part of randomly initialized weights yet to be trained on the mice videodata. The final model takes 9 stacked frames that form one video as input and predict an action for the middle frame.

Pre-processing The original dataset was not pre-processed, albeit every input to the model was pre-processed individually for every prediction. For each input video, 27 frames were uniformly subsampled by steps of 3 to form 9 consecutive frames that would be the small video as input to the model. To better generalize to different camera angles and types, and video distortions, the input was subjected to random small affine transformations (rotation, shearing and translation) with probability 0.8 and random uniform noise with probability 0.5 for each batch during training. This noise was always pixel-wise applied in the uniform range is -1.5 to 1.5 times the standard deviation per pixel over the batch. Finally, for validation there was no batch-wise distortion to stay true to the original data.

Metrics To quantify success of the network in predicting object exploration several metrics were calculated per epoch for testing performance on the training and validation set. The first metric, which was also the loss function, is the binary cross-entropy loss, calculated as described in section 1.1.1. The second metric is the area under the receiver operating characteristics curve (AUROC), calculated as:

$$AUROC = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}_{p_i > p_j}$$

This runs over all m datapoints with label 1, and all n datapoints with label 0. Here, p_i and p_j are the probabilities of datapoints i and j assigned by the network. The indicator function $\mathbf{1}$ outputs 1 iff the condition $p_i > p_j$ is satisfied.

Training The network was trained for 139 epochs and 1000 iterations per epoch on the training data set. Each iterations had a batch size of 7 videos, such that the size of the input was $7 \times 9 \times 384 \times 512 \times 3$ with the last dimension being colour. The loss per batch was calculated as the binary Cross-Entropy loss function and weights were optimized using stochastic gradient descent.

2.2.2 Wall Exploration and Corner Sitting

The goal is to find some mapping per $\mathcal{V}$ such that the output is a time-series vector $\boldsymbol{x} = \begin{pmatrix} x_1 & \dots & x_t \end{pmatrix}$ of length t that per frame i outputs what action the mouse is doing x_i, with action set:

$$\chi = \{\text{exploring wall}, \text{sitting in corner}\}$$

This is a function $f_{WC} :: \left\{ \mathcal{V} \mapsto \boldsymbol{x} = \begin{pmatrix} x_1 & \dots & x_t \end{pmatrix} \right.$.

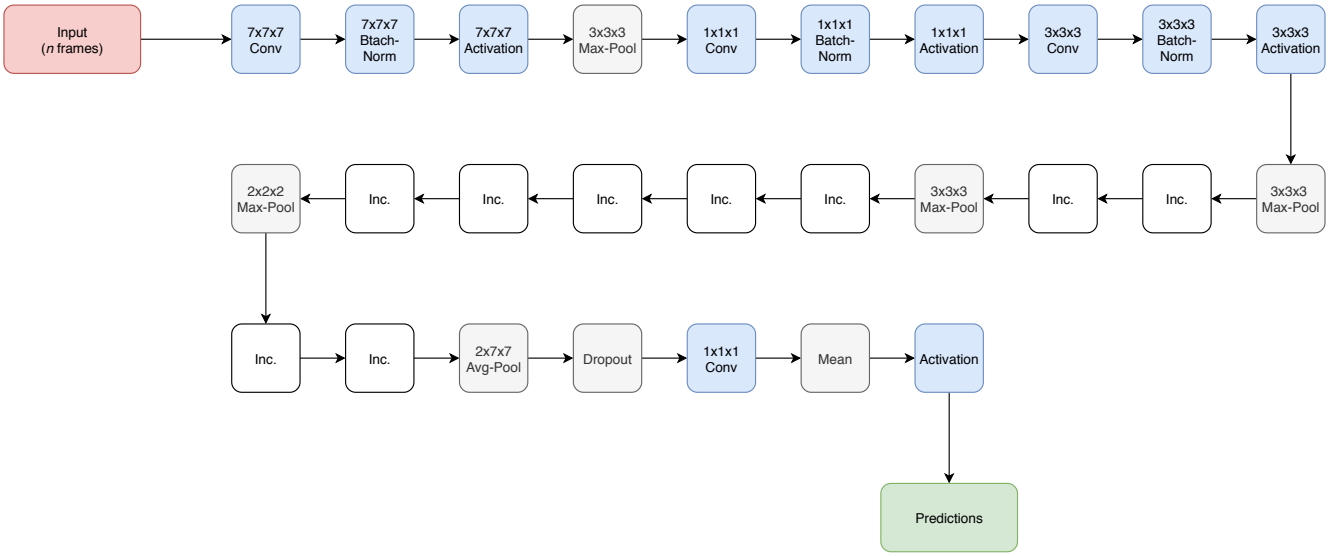
Dataset A dataset was created of 8000 images (width: 384, height: 512) of mice in either the the green or white background Object-Space task box. These images were manually labeled, using DeepLabCut (Nath* et al., 2018), for four body parts: 1. Nose, 2. Right ear, 3. Left ear, 4. Back, and 5. Tail base.

Pose Estimation An ANN was trained on 90% of the dataset and validated on 10% of the dataset, using DeepLabCut (Nath* et al., 2018). This model extracted four bod parts of a mouse in a single image: 1. Nose, 2. Right ear, 3. Left ear, 4. Back, and 5. Tail base. Using the pose, the head direction (HD) of the mice can be calculated as the angle between the mean vector of the body parts $\lambda = \{$Right ear, Left ear, Back$\}$ pointing to the Nose; more formally $HD = arctan2(y, x)$, with (x, y) being coordinates calculated in its own plane calculated as $(x, y) = \frac{1}{3} \sum_{bp \in \lambda} (x, y)_{bp} - (x, y)_{nose}$. The result is a model that can estimate both the pose and head direction of a mouse in a single image of the Object-Space task.

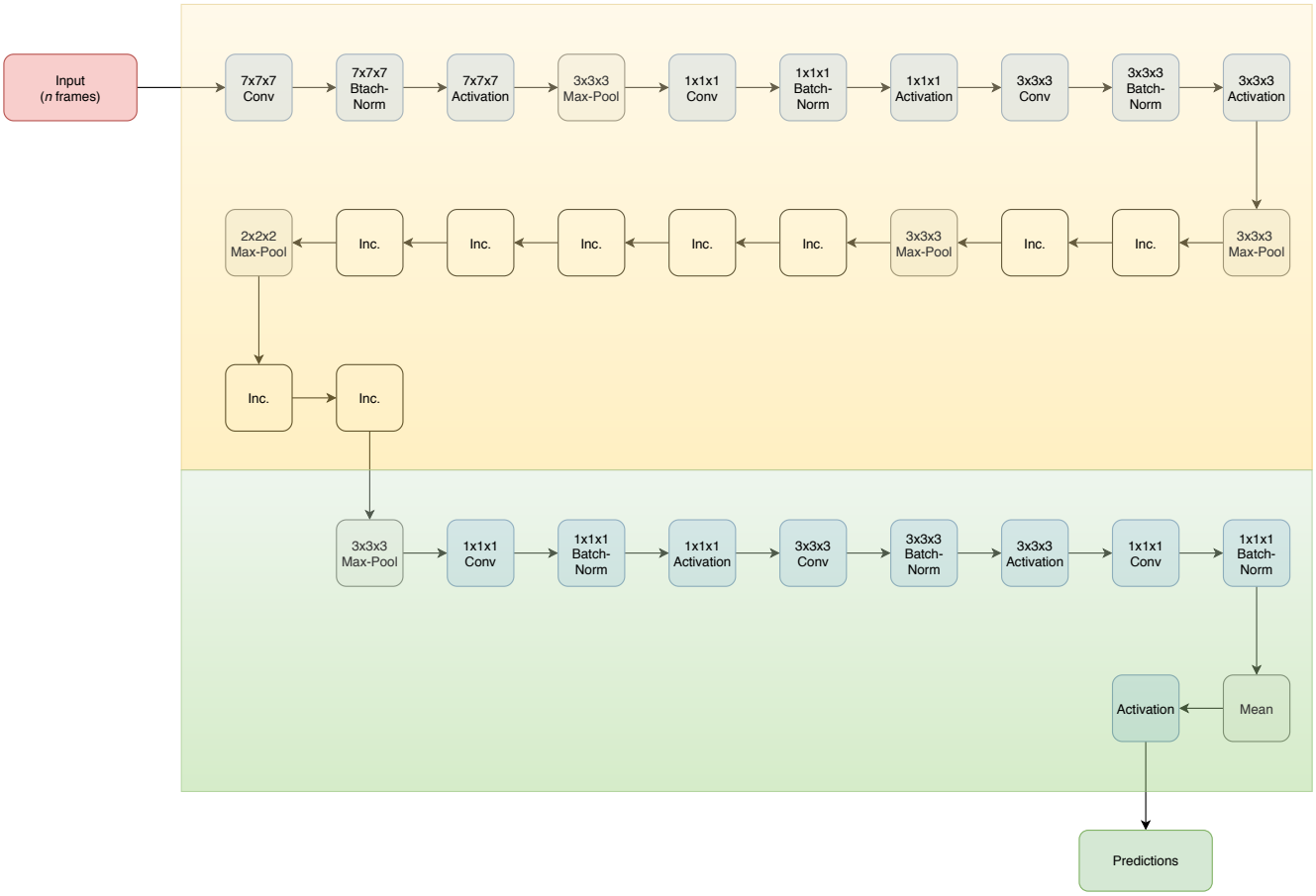
Box Template To relate mouse pose to any area-related behaviour, box templates were made where the wall and corner areas are represented as rectangles and circles respectively. For walls this means that each cardinal direction wall is represented by a rectangle spanning the wall, including a bit of space next to it. For corners this means that each corner is represented by a circle, with its origin in the most outer part. One example template can be found in Figure 2.3.

The Model A rule-based model was made that takes mouse pose and box template as input and outputs the set of actions that is applicable to that frame. The mouse's location is represented as the mean xy-coordinate of the Nose, Left ear, and Right Ear: $(x, y) = \frac{1}{3} \sum_{bp \in \lambda} (x, y)_{bp}$. To evaluate whether the mouse is in the corner, the mouse's location can be checked to be in one of the circles that represent this area. To evaluate whether the mouse is exploring the wall, the mouse's location and head direction can be checked per wall. For example, if the mouse is in the north wall's area then the head direction must be $HD \in [0.1 \cdot \pi, 0.9 \cdot \pi]$. Here the constants 0.1 and 0.9 represent a 90% field of view for the mouse in π distance.

The final function that maps some video to the set of actions per frame $f :: \left\{ \mathcal{V} \mapsto \boldsymbol{x} = \begin{pmatrix} x_1 & \dots & x_t \end{pmatrix} \right.$ can be constructed by concatenating the outputs of both f_E and f_{WC} (see Figure 2.4).



(A) The original inflated-inception-V1 network structure for human action recognition by Carreira and Zisserman (2017).



(B) Modified network structure of subfigure (A). Here, part of the network in the orange box is copied from the original network (including weights). The top part (following the last Inc. layer) of the old network is cut off and replaced by a the new structure in the green box (random initial weights).

FIGURE 2.2: The ANN structure used for kinetic action recognition, where subfigure (A) is the original network used for human action recognition and subfigure (B) is the re-structured network for mice action recognition. Inc. is an inception module as described in Figure 1.1. The first input is always n stacked frames that form a small video, and the output is predicted probability of all actions. Instead of probability for 400 human action classes, the predictions are now two-class: exploring object and ¬exploring object.

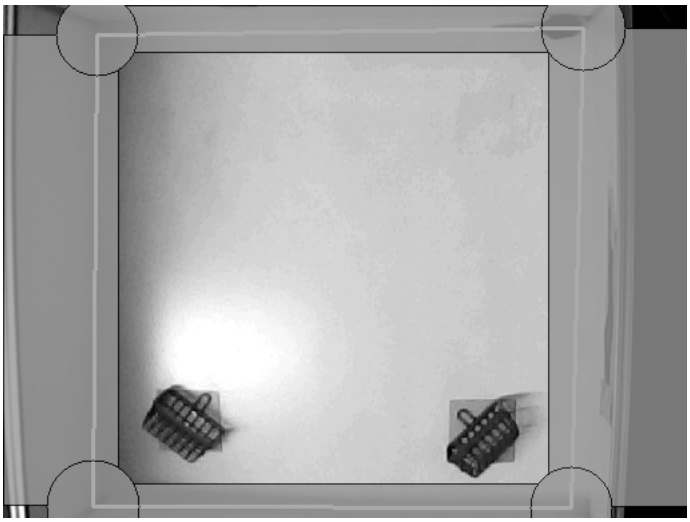


FIGURE 2.3: Example template of a box as seen by the rule-based model. Rectangles represent wall areas and circles represent corner areas.

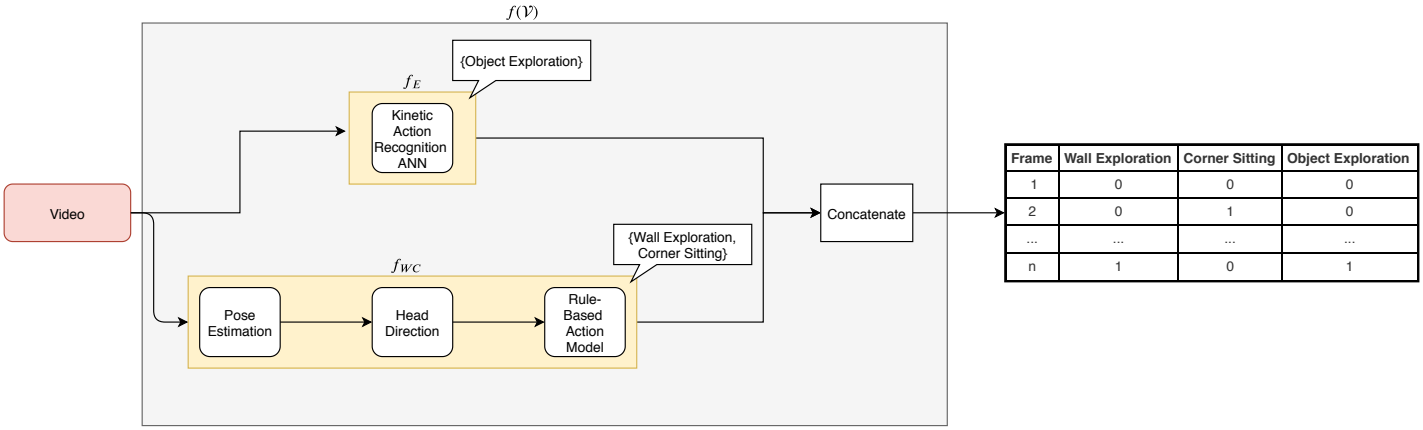


FIGURE 2.4: The full model for the automatic behavioural scoring $f(\mathcal{V})$. First, the video is passed along both the Kinetic Action Recognition ANN f_E and the rule-based model f_{WC} separately. These models output {Object Exploration} and {Wall Exploration, Corner Sitting} respectively per frame. These actions are then concatenated into a final judgement of all actions per frame of the video.

2.3 Genotyping

The methods used to predict the genotype of each mouse subject per trial in a certain condition (i.e. video) will be described in this section.

The data after scoring videos $\mathcal{V}$ in the Object-Space task consists of approximately $|Subjects| \times |Trials| \times |Condition| = |Subjects| \times 21 \times 3$ sequences $\boldsymbol{x} = \begin{pmatrix} x_1 & \dots & x_t \end{pmatrix}$, where x_i is either exploring object 1, exploring object 2 or not exploring at time i. The sequences $\boldsymbol{x}$ were extracted using the automatic behavioural scoring model described in Section 2.2. The goal here is to find some mapping per $\boldsymbol{x}$ such that the output is a boolean $\mathcal{B}$ describing the subject to be either an ehmt1$^{+/+}$ (0) or ehmt1$^{+/1-}$ (1) genotype, that is a function $g :: \begin{Bmatrix} \boldsymbol{x} \mapsto \mathcal{B} \end{Bmatrix}$.

2.3.1 Features

To predict predict genotype for each sequence $\boldsymbol{x}$, features that potentially explain behavioural variance as a result of genotype were extracted. An overview of all features can be found in Table 2.1.

The first feature is *First object*, which is set to 1 if object 1 was first explored and set to 2 if object 2 was first explored in the trial. Secondly, *First object latency* was calculated was the total time (i.e. in frames or time) that has passed until the mouse first explored any object. The next two features, $Stay_1$ and $Stay_2$, describe the likelihood over the whole trial of transitioning to exploring object i after object i was just explored. This likelihood satisfies the Markov property, which refers to that the next state (i.e. exploring object 1 or 2) depends only on the current state. Specifically, the feature $Stay_i$ was calculated as follows: $p(E_k = i | E_{k-1} = i) = \frac{\sum_{k=2}^{E} \delta_{ii}^{k} \cdot \delta_{ii}^{k-1}}{\sum_{k=2}^{E} \delta_{ii}^{k-1}}$, which iterates over all object explorations E to evaluate the proportion of specific object i explorations k that followed from an exploration of object i for $k-1$. Here, δ is the dirac function which is set to 0 if its subindices are identical, and 0 otherwise. The next feature SS_1 (steady state probability for object 1) is calculated based on the transition matrix $\mathcal{P} = \begin{pmatrix} Stay_1 & 1 - Stay_1 \\ 1 - Stay_2 & Stay_2 \end{pmatrix}$. It is calculated as the first element of μ in $\mu\mathcal{P} = \mu$, meaning that the probability of exploring object 1 next does no longer change after exploring any next object.

Next, the *Perseverance Index* is an index ranging from -1 to 1, where -1 represents a tendency to switch objects between explorations, 1 represents a tendency to re-explore the same object after it was just explored, and 0 represents no tendency. Specifically, the feature was calculated was follows: $PI = \frac{\sum_i stay_i - (\sum_i 1 - stay_i)}{\sum_i stay_i + (\sum_i 1 - stay_i)}$, where $stay_i$ is the feature *stay* for object i. The next feature, *n_transitions* is the total number of transitions made between any objects during the whole trial and is in essence the total number of object explorations minus one. The next five features are min_i*_n_explore*, where i is in the range from 1 to 5. This describes the cumulative number of object explorations per minute for the first five minutes. In line with the previous feature, the next ten features are min_i*_n_explore_object*$_j$, which for each object describes the cumulative number of explorations up until that minute.

The next five features are min_i*_time*. where i is is in the range from 1 to 5. This describes the total cumulative time of exploring any object up until minute i. The next ten features are min_i*_object*$_j$*_time*, where j is either object 1 or 2. This describes the total time a specific object was explored up until that minute minute.

The next feature is *bout_time*. This feature describes the mean specific object exploration (bout) time over the whole trial. Thus, how long a mouse keeps exploring a single object. It is calculated as $\frac{\text{min}_5\text{\_time}}{\text{min}_5\text{\_n\_explore}}$. The next two features bout_time_obj$_j$ are in line with the previous one. It describes the bout time for each specific object over the whole trial. It is calculated as $\frac{\text{min}_5\text{\_object}_j\text{\_time}}{\text{min}_5\text{\_n\_explore\_object}_j}$.

Lastly, the five features min_i*_DI* represent the Discrimination Index (DI) for each minute up until that minute. The DI is an index ranging from -1 to 1, where -1 represents a preference for exploring object 2, 1 represents a preference for object 1, and 0 represents no preference. Specifically, the feature

was calculates a follows: $\frac{(\sum_k^E \delta_{1i}^k \cdot time_k) - ((\sum_k^E \delta_{2i}^k \cdot time_k)}{\sum_k^E time_k}$, were $time_k$ is duration of a specific exploration k, and E is all explorations up until that minute.

Feature	Description	Calculation
First object	Whether the first object explored was either object 1 or object 2	
First object latency	Latency to first exploration of any object	
Stay$_i$	Likelihood of exploring object i next, after having last explored object i. Where i is either 1 or 2.	$\frac{\sum_{k=2}^E \delta_{ii}^k \cdot \delta_{ii}^{k-1}}{\sum_{k=2}^E \delta_{ii}^{k-1}}$, where k is a specific exploration any of object and δ is the dirac function
SteadyState$_1$(SS_1)	The probability of exploring object 1 after many explorations	First element of μ, in $\mu\mathcal{P} = \mu$. With transition matrix $\mathcal{P}$
Perseverance Index	Index ranging from -1 to 1 represent the tendency of switching between objects during exploration as -1, the tendency to re-explore the same object during exploration as 1, and no tendency as 0	$\frac{\sum_i stay_i - (\sum_i 1 - stay_i)}{\sum_i stay_i + (\sum_i 1 - stay_i)}$, where $stay_i$ is the previous feature for object i
n_transitions	Total number of transitions made between objects during the whole trial.	
min$_i$_n_explore	Per minute i, the total number of explorations of any object up until that minute	
min$_i$_n_explore_object$_j$	Per minute i, the total number of explorations of object j up until that minute	
min$_i$_time	Per minute i, the total time any object was explored up until that minute	
min$_i$_object$_j$_time	Per minute i and object j, the total time that object was explored up until that minute	
bout_time	Mean time of explorations of any object	$\frac{\text{min}_5\text{\_time}}{\text{min}_5\text{\_n\_explore}}$
bout_time_obj$_j$	Mean time of explorations of object j	$\frac{\text{min}_5\text{\_object}_j\text{\_time}}{\text{min}_5\text{\_n\_explore\_object}_j}$
min$_i$_DI	Per minute i, the the discrimination index (DI) up until that time representing a preference for exploring object 2 as -1, a preference for exploring object 1 as 1, and no preference as 0	$\frac{(\sum_k^E \delta_{1i}^k \cdot time_k) - ((\sum_k^E \delta_{2i}^k \cdot time_k)}{\sum_k^E time_k}$, were $time_k$ is duration of a specific exploration k, and E is all explorations up until that minute

TABLE 2.1: Features extracted from the sequence data per trial. Each feature is described, and, where applicable, the calculation is shown. In total 45 features were calculated. All features were calculated only for the first five minutes of a trial.

2.3.2 Implementation

Dataset After feature engineering, the dataset consisted of 1738 datapoints (i.e. rows). Each datapoint entails information about the trial such as condition, trial number, subject number, and subject genotype. Next to this, each datapoint include all the features corresponding to its trial (i.e. video). In this dataset, only features are considered predictors and the genotype is the target variable. All analyses in this section will be done on the dataset split by the three conditions (stable, overlapping, random) and all conditions pooled, such that results may be compared between conditions. This resulted in a total of 1738 (ehmt1$^{+/+}$: 1213, ehmt1$^{+/-}$: 525) datapoints for the pooled trials, 590 (ehmt1$^{+/+}$: 416, ehmt1$^{+/-}$: 174) datapoints for the overlapping trials, 585 (ehmt1$^{+/+}$: 409, ehmt1$^{+/-}$: 176) datapoints for the stable trials, and 563 (ehmt1$^{+/+}$: 388, ehmt1$^{+/-}$: 175) datapoints for the random (control) trials.

Pipeline A diagram of the logic to classify genotype based on the features in the dataset can be found in Figure 2.5. The figure depicts that first the dataset is randomly split into a training (90%) and validation (10%) set. Then, the features are selected as predictors which are used as input for multiple independent classifiers (see next paragraph). . The parameters for each classifier are then optimized using 10-fold cross-validation grid search with AUROC as the performance metric (Kohavi, 1995; Bergstra and Bengio, 2012). After optimal parameters for each classifier is set, this results in models that can now predict genotype for each trial using the given features. The models are then tested against the final validation set with AUROC as a metric.

Models

Random Forest The Random Forest classifier is a bootstrapping algorithm that parallelly constructs multiple decision trees that are subsequently merged together to make a final prediction (Breiman, 2001). To clarify, the decision trees learn how to split the dataset into smaller subsets in order to predict the outcome. Such a decision tree consists of nodes (i.e. condition) and edges (i.e. decisions), which are build in layers until the decision tree is optimized by measure of some impurity criterion. The randomness in the algorithm is that, while splitting nodes, the best feature is chosen among a random subset of features. This randomness creates a variable forest of decision trees, consequently reducing overfitting. The following parameters were optimized: maximum number of features used for each node split, minimum number of leafs to split an internal node, and criterion measure (gini impurity, entropy).

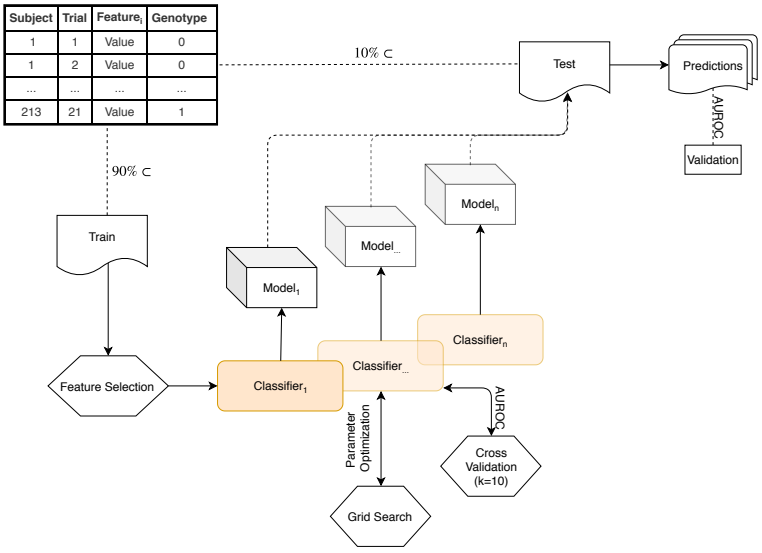


FIGURE 2.5: This figure depicts that first the dataset (upper left) were split into a training ($\subset$ 90%) and validation ($\subset$ 10%) set. Using the training set, all classifiers are trained and optimized for its respective parameters using grid search with 10-fold cross validation. A classifier with its parameters optimized for the dataset is called a model (for specific models, see next section). Each model then predicts genotype on the validation set and is tested on its AUROC score.

XGBoost The XGBoost classifier is a gradient boosting algorithm that sequentially constructs many weak learners (i.e. shallow decision trees) that are subsequently aggregated to make a final decision (Chen and Guestrin, 2016). Whereas

the Random Forest model merges the outcome of many (deep) decision trees in order to reduce model variance, XGBoost recursively builds many shallow decision trees that uses the residuals of its predecessor, which ultimately reduces bias. To clarify, first a base-model $F_1(x) = y$ is fit to the data upon which a residual-model is fit to its residuals such that $h_1 = y - F_1(x)$, then $F_i(x) = F_{i-1}(x) + h_{i-1}(x)$, where y is the target value, F_i is the current model, and h_i is the model on the residuals of F_i. This procedure can go on until either the model has no error or a maximum number of M models has been reached. The following parameters were optimized: learning rate, maximum tree depth, gamma (node split regularization), and minimum sum of instance weight (hessian) needed in a node child.

Feature Importance For all models, feature importance's were calculated as relative drop-column importance. That is, for each feature $f \in \mathcal{F}$ in base-model $\mathcal{M}_{\mathcal{F}}(\boldsymbol{x})$, the model is re-trained on the training data without that feature such that features $\mathcal{F}' = \mathcal{F} \setminus \{f\}$ and the model without that feature is $\mathcal{M}_{\mathcal{F}'}(\boldsymbol{x})$. Then the absolute feature importance is calculated as contribution to accuracy: $\mathcal{I}_f = AUROC(\mathcal{M}_{\mathcal{F}}(\boldsymbol{x})) - AUROC(\mathcal{M}_{\mathcal{F}'}(\boldsymbol{x}))$. This is then normalized to be between 0 and 1 and summing to 1 by: $\mathcal{I}_f = \frac{\mathcal{I}_f - \underset{\mathcal{I}_{\mathcal{F}}}{\text{argmin}}}{\sum_f (\mathcal{I}_f - \underset{\mathcal{I}_{\mathcal{F}}}{\text{argmin}})}$. Thus, the feature importances represent an importance ranking for each feature within a model. Furthermore, feature importances do represent the importance of that feature plus all possible interactions with all other features.

Data Analysis To test whether the AUROC of each model was above the expected value of chance, they were tested under the permutation distribution (Welch, 1990; Ernst, 2004). That is, the classification of genotype procedure was repeated (simulated) 1000 times with the same model and the same features with its labels randomly permuted. The probability of getting the original model's AUROC is then given by the percentage of simulations that had an equal or higher AUROC than the original model.

To gain further insight in genetic differences within feature scores, the top 5 features per model were tested under the permutation distribution. That is, genotype (0, 1) was treated as a between subject-factor for each feature for a two-sided test. Specifically, for each feature the group mean differences were calculated 10000 with the genotype labels permuted. Then, the probability of either genotype being high than the other is given by the percentage of simulation where the mean difference is equal to or higher/lower than the original group mean difference.

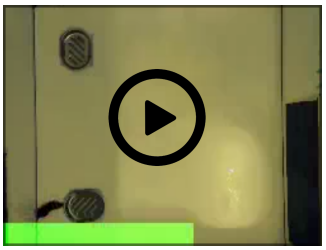
Results

3.1 Automatic Behavioural Scoring

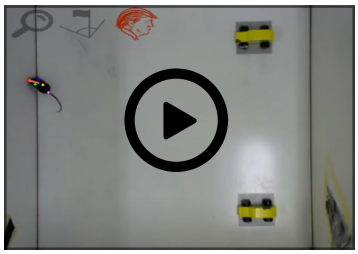
Object Exploration Model The binary Cross-Entropy loss and AUROC over epochs of the deep inflated inception kinetic action recognition ANN are depicted in Figure 3.2a. Both the training and validation sets' loss decrease over-time, with the training set (loss: 0.136) having a slightly lower loss than the validation set (loss: 0.152). Furthermore, the AUROCs of both the validation and training increase over epochs, with the training set (AUROC: 0.9914) having a slightly higher AUROC than the validation set (AUROC: 0.9852). Next, the ROC-curve of the last epoch is shown in Figure 3.2b. The network has both high sensitivity and specificity for both the training and validation set.

Finally, since this model works with video data, its performance is perhaps best illustrated by showing predictions on actual footage. The video is shown in Figure 3.1a. These predictions are made on a video that the network is neither trained on nor validated with. The changing length and color line at the bottom of the video depicts the probability of the mouse exploring an object that the network ascribes to the frame. This bar is red when $p < 0.5$ and green otherwise. Recall that for the final model that includes all actions, this probability is binary at a threshold of 0.5.

Final Model The final model predicts all actions $\chi = \{\text{exploring object}, \text{exploring wall}, \text{sitting in corner}\}$ per frame of a video $\mathcal{V}$. The final model is a concatenation of the exploration model and the rule-based model, to respectively predict $\{\text{Object Exploration}\}$ the remaining actions $\{\text{Wall Exploration}, \text{Corner Sitting}\}$. Since the latter part of the model has no ground truth labeled data, it is best validated by qualitatively inspecting its results. A video with predictions is shown in Figure 3.1b. These predictions are made on video data that has not been used in any training in any part of either model.

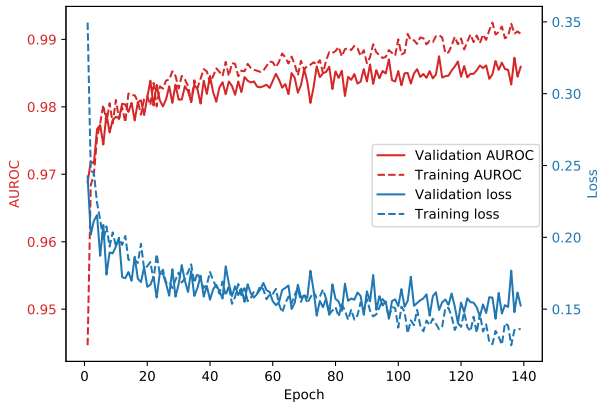


(A) A video of predictions made on mouse Object Exploration by the model in Figure 2.2b. The bar on the bottom represents the probability p that the model ascribes to the mouse exploring the object. This bar is red when $p < 0.5$ and green otherwise.

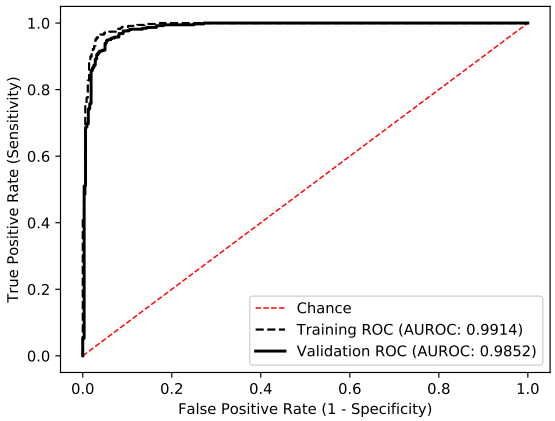


(B) A video of predictions made on mouse Object Exploration, Wall Exploration, and Corner Sitting by the model in Figure 2.4. Here, the icons in the upper left corner represent the actions and are black when no action is predicted. Object Exploration is indicated by the magnifying glass turning blue, Corner Sitting is indicated by the flagged corner turning green, and Wall Exploration is indicated by the Donald Trump face turning orange. Furthermore, coloured dots are projected onto the mouse's limbs and an arrow is projected to indicate the mouse's head direction.

FIGURE 3.1: Videos of predictions by both Automatic Behavioural Scoring models. Please click on the respective image to play the video.



(A) Loss and AUROC by the kinetic action recognition ANN for both the training and validation set per epoch. The loss decreases over epochs and is lower for the training than the validation set. The AUROC increases over epochs and is higher for the training than the validation set.



(B) Receiver Operating Characteristics curve for the last epoch in the kinetic action recognition ANN for both the training and validation set. The curve shows the model to be sensitive and specific for both the training and validation set.

FIGURE 3.2: Statistics for the kinetic action recognition ANN that predicts Object Exploration of mice. To see the model in action, the reader is referred to this video.

3.2 Genotyping

Model Performances Performance of each model for each condition (all pooled, stable, overlapping, random) are depicted in Figure 3.3a with its accompanying ROC-curves in Figure 3.3b. All models except the ones in the control condition have an AUROC > 0.61 that is significant ($\forall i : p_i < 0.001$) under the permutation distribution. Furthermore, inspecting each significant model's ROC-curve shows that all models have relatively high specificity relative to their sensitivity.

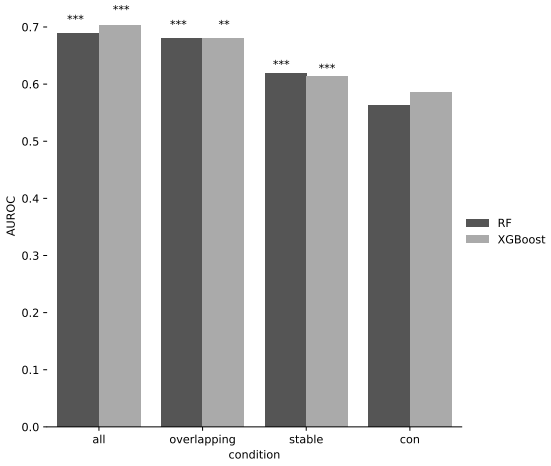
For the models trained on all conditions pooled, both the RF (AUROC ≈ 0.69, $p < .001$) and XGBoost (AUROC ≈ 0.7, $p < .001$) are predicting genotype of mice based on single trials with a performance above chance level. Furthermore, both the RF (sensitivity≈ 0.49, specificity≈ 0.89) and XGBoost (sensitivity≈ 0.51, specificity≈ 0.89) models have higher specificity than sensitivity. This means that the model's are very specific in ascribing the ehmt1$^{+/+}$ genotype on the potential cost of missing cases that should have been detected.

For the models trained on trials in the stable condition, both the RF (AUROC ≈ 0.62, $p < .001$) and XGBoost (AUROC ≈ 0.61, $p < .001$) are predicting genotype of mice based on single trials with a performance above chance level. Furthermore, both the RF (sensitivity≈ 0.35, specificity≈ 0.88) and XGBoost (sensitivity≈ 0.35, specificity≈ 0.87) models have higher specificity than sensitivity.

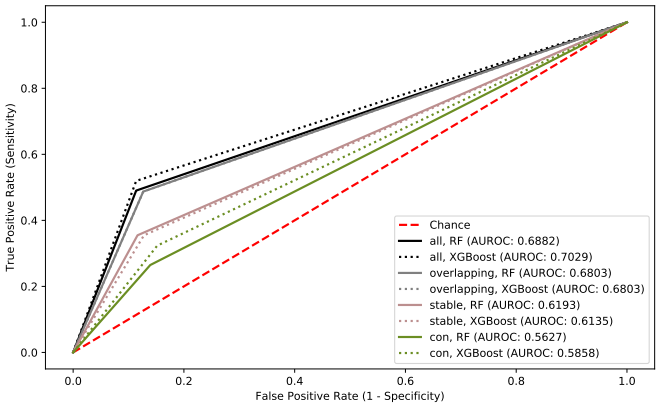
For the models trained on trials in the overlapping condition, both the RF (AUROC ≈ 0.68, $p < .001$) and XGBoost (AUROC ≈ 0.68, $p < .001$) are predicting genotype of mice based on single trials with a performance above chance level. Furthermore, both the RF (sensitivity≈ 0.48, specificity≈ 0.87) and XGBoost (sensitivity≈ 0.48, specificity≈ 0.87) models have higher specificity than sensitivity.

For the models trained on trials in the random condition, both the RF (AUROC ≈ 0.56, $0.05 < p < 0.07$) and XGBoost (AUROC ≈ 0.59, $0.05 < p < 0.0$) are not significantly predicting genotype of mice based on single trials with a performance above chance level.

However, the same trend shows as with the aformentioned conditions that both the RF (sensitivity≈ 0.26, specificity≈ 0.86) and XGBoost (sensitivity≈ 0.32, specificity≈ 0.85) models have higher specificity than sensitivity. This The insignificant performance's seems to be the consequence of a low sensitivity.



(A) AUROCs by all classifiers for all subsets of data (all conditions, stable, overlapping, random). Asterisks represent the p-value of its respective AUROC under the permutation distribution: * = ($p < .05$), ** = ($p < .01$), *** = ($p < .001$).



(B) Receiver Operating Characteristics curves for all classifiers and all subsets of data.

FIGURE 3.3: Statistics for the genotype classifiers based on the features by model labeled data for all conditions in the Object-Space task.

Feature Importances Top ten features' relative importances and top five feature mean differences were tested under the permutation distribution. A high relative feature importance means that a model uses that specific feature and all its potential interactions with other features. Relating this to the feature differences between genotype, this implies that a significant difference of a feature between genotype might be used directly in a model, whereas an insignificant difference might be used in interaction with other features. This complexity of features in decision trees is illustrated in Figure 3.4 for the overlapping condition. This particular decision tree uses an interaction between *min2_obj1_time* and *min4_DI* to split the initial major parts of the data for classification.

For the models trained on all conditions pooled, top ten features per model are depicted in Figure 3.5a. The mean differences of a feature between genotype of the top five feature per model are depicted in Figure 3.6. All features, except *bout_obj2_time*, *min2_n_explore* and *stay1*, differ significantly ($p < 0.05$) in their mean.

For the models trained on trials in the stable condition, top ten features per model are depicted in Figure 3.5b. The mean differences of a feature between genotype of the top five feature per model are depicted in Figure 3.6. Six features differ significantly ($p < 0.001$): *min2_n_explore_time*, *min3_explore_time*, *min_obj2_time*, *min5_explore_time*, and *min5_obj1_time*. With all cases the mean being higher for the ehmt1$^{+/+}$ than the ehmt1$^{+/-}$ genotype.

For the models trained on trials in the overlapping condition, top ten features per model are depicted in Figure 3.5c. The mean differences of a feature between genotype of the top five feature per model are depicted in Figure 3.6. Almost all features differ significantly ($p < 0.05$) in their mean, except *bout_time* and *min2_n_explore*. Notably, this is the only model that includes multiple DI features that all differ significantly in their mean with the ehmt1$^{+/-}$ genotype consistently having a higher DI than the ehmt1$^{+/+}$ genotype.

For the models trained on trials in the random condition, top ten features per model are depicted in Figure 3.5d. The mean differences of a feature between genotype of the top five feature per model are depicted in Figure 3.6. Only two features differ significantly in their mean: *min4_n_explore* ($p < 0.001$) and *min5_explore_time* ($p < 0.001$), where in both cases the ehmt1$^{+/+}$ genotype has the highest explore times.

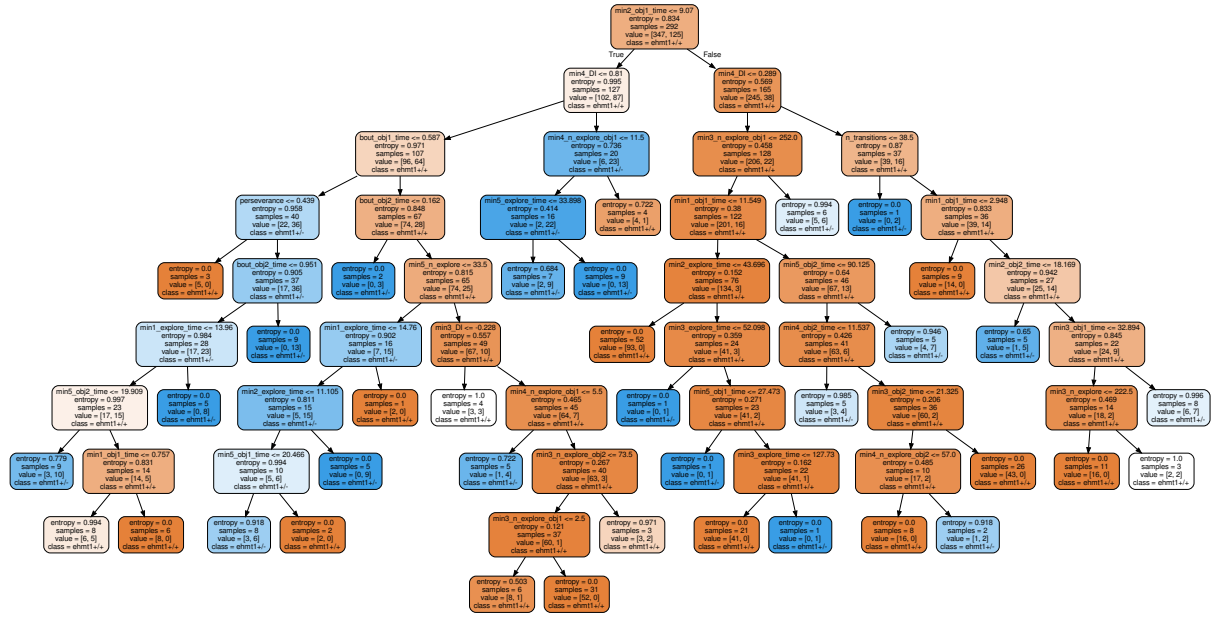
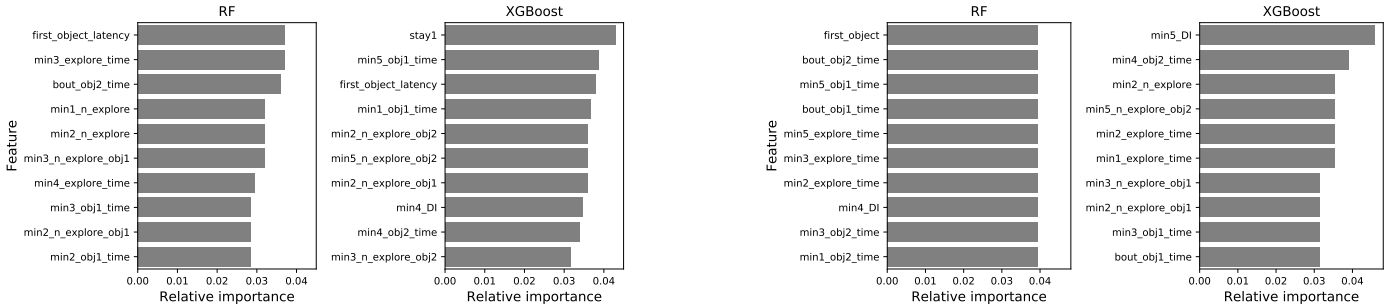
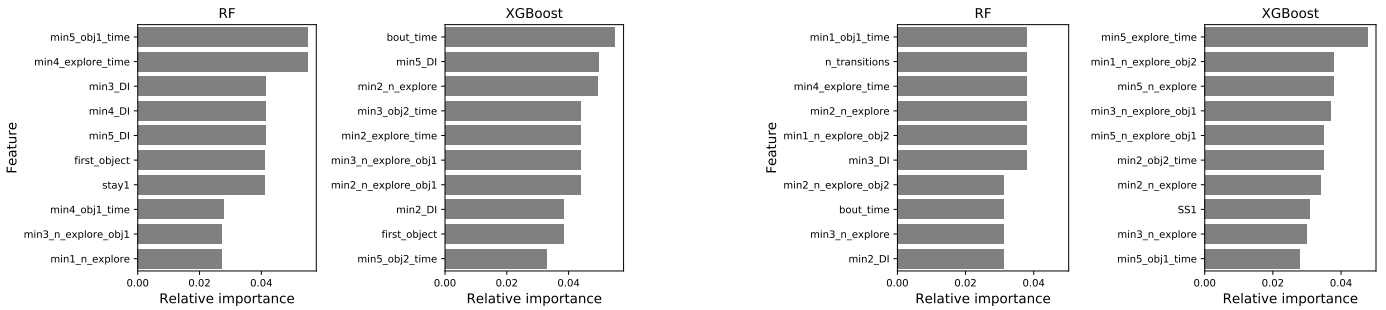


FIGURE 3.4: Example decision tree for models based on data from the overlapping condition trials. This particular decision tree seems to use an interaction between *min2_obj1_time* and *min4_DI* to split the first major part of the data into classes. This shows that classification is no simple single variable cut-off decision, but a complex process. Note that one decision tree is not fully representative for all the trees used in the ensemble.



(A) Top features for each model using the trials in all conditions.

(B) Top features for each model using the trials in the stable condition.



(C) Top features for each model using the trials in the overlapping condition.

(D) Top features for each model using the trials in the control condition.

FIGURE 3.5: Top 10 features and their relative feature importance per model for all subsets of data (all conditions, stable, overlapping, random).

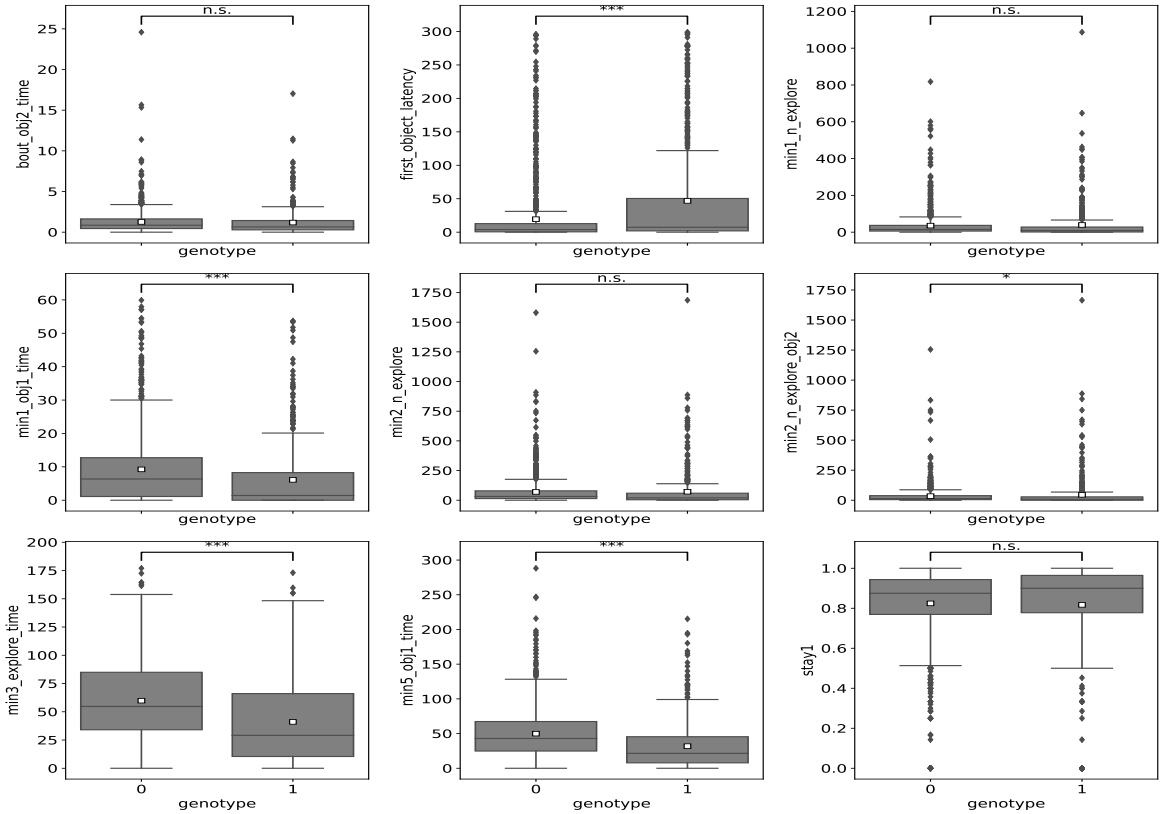


FIGURE 3.6: (A). Top features and their box plot per genotype for each model using the trials in all conditions.

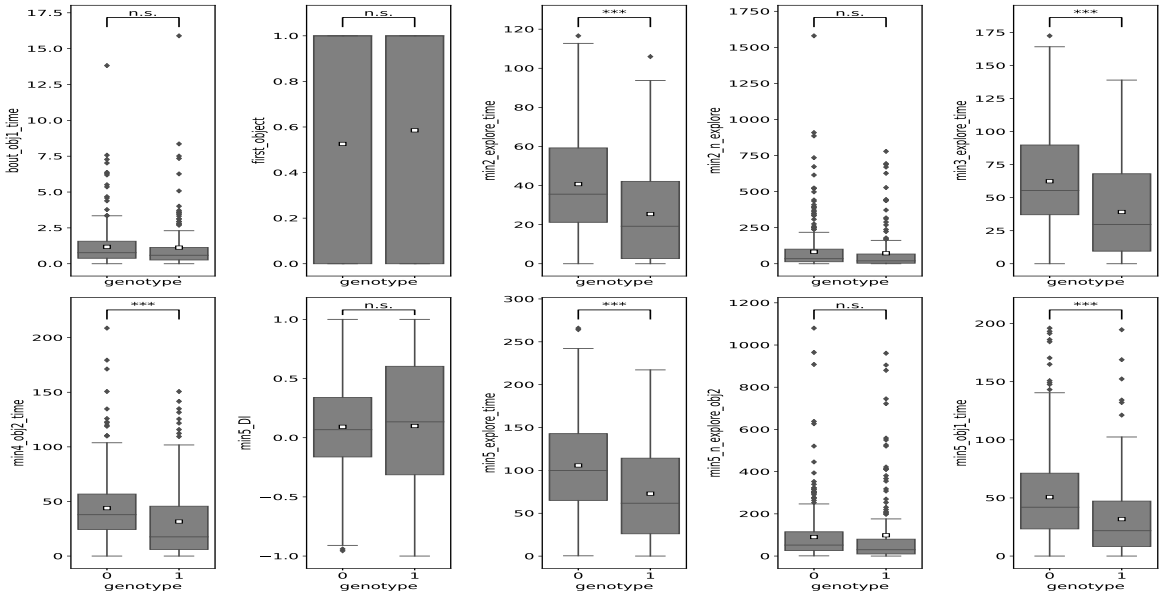


FIGURE 3.6: (B). Top features and their box plot per genotype for each model using the trials in the stable condition.

Box plots of the top 5 features per model for all subsets of data (all conditions, stable, overlapping, random). The white square represents the mean. The difference of the mean between genotype (0,1) was tested for each feature under the permutation distribution. Asterisks represent the p-value of the respective feature differences: * = ($p < .05$), ** = ($p < .01$), *** = ($p < .001$), n.s. = ($p > 0.05$).

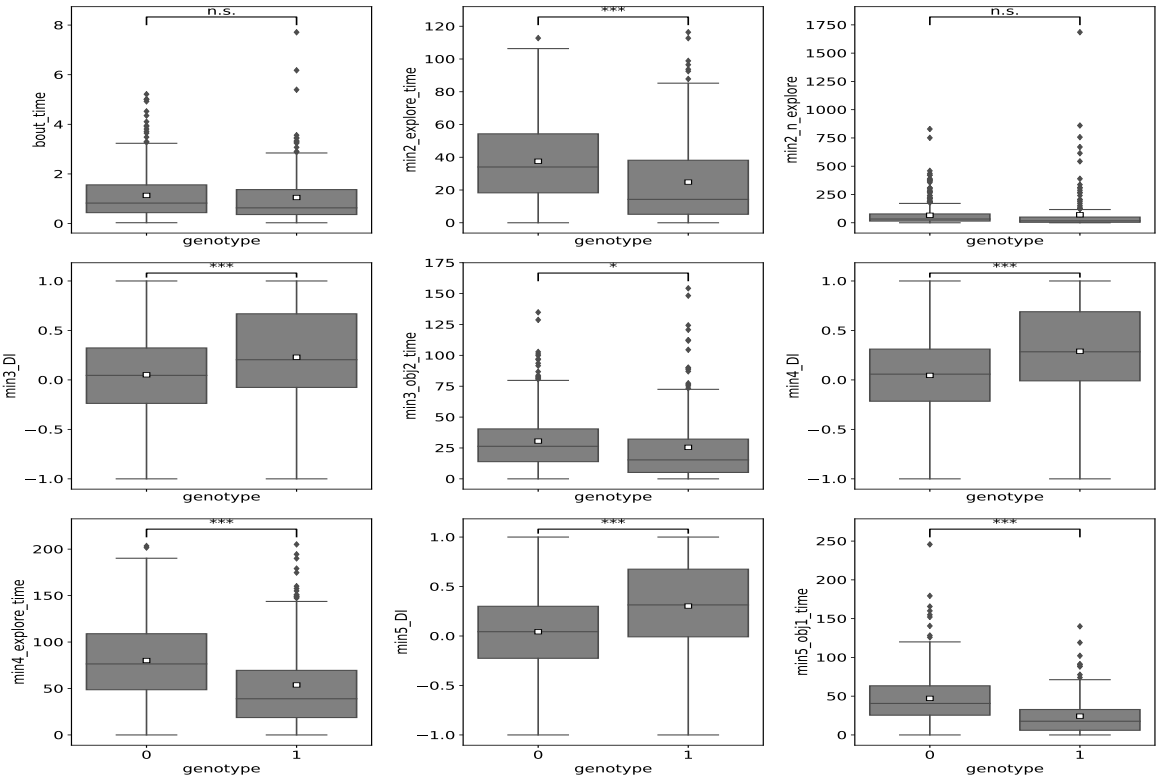


FIGURE 3.6: (C). Top features and their box plot per genotype for each model using the trials in the overlapping condition.

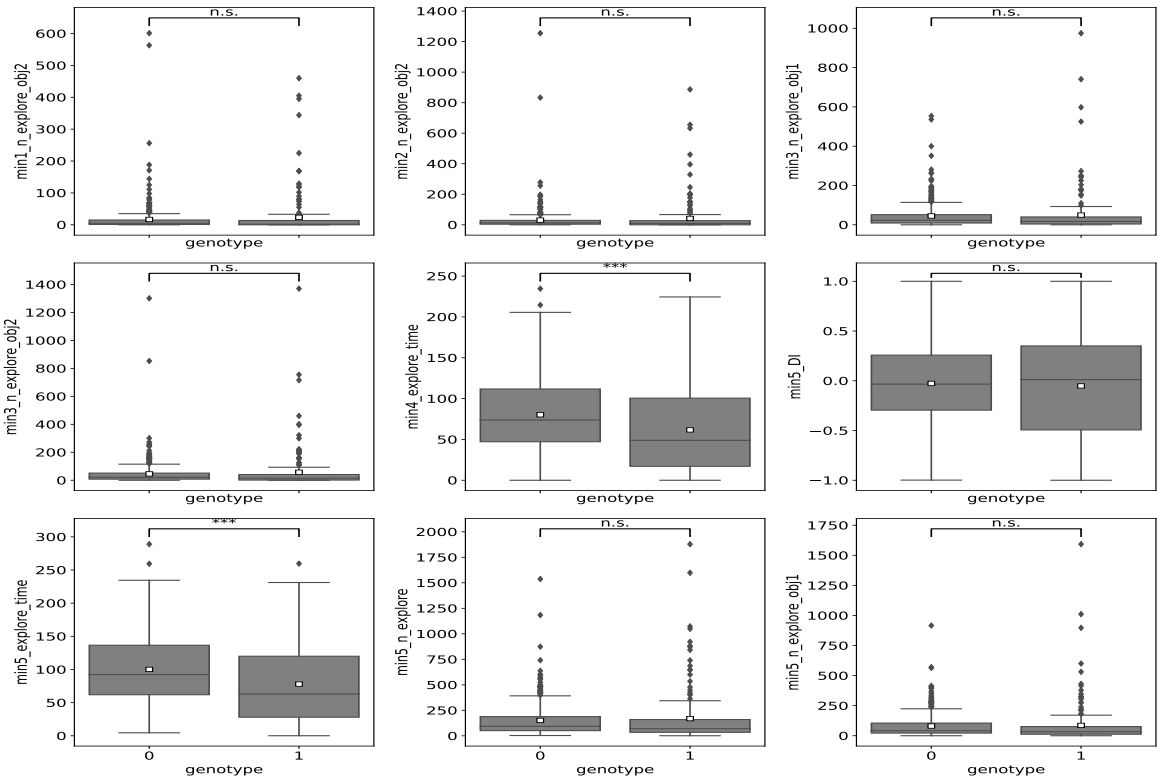


FIGURE 3.6: (D). Top features and their box plot per genotype for each model using the trials in the random condition.

Discussion

It has been a long standing challenge in translational neuroscience to reliably extract meaningful information from rodents performing composed modules of behaviour. One major part of this challenge is due to the unreliability and laboriousness of human visual inspection of large amounts of video data (Blanchard, Griebel, and Blanchard, 2003; Pandey et al., 2008). This problem may be tackled by applying advances machine learning to track rodents and identifying their behaviours. Previous attempts at categorizing behaviour at the spatiotemporal scale have either focused on latent expressions or fully predefined models (De Chaumont et al., 2012; Wiltschko et al., 2015). This work presented a method to score sequences of varying behaviours in mice roaming a square box. By applying pose estimation and optic flow techniques from the machine learning field, it has been shown that it is possible to perform a fully automated behavioural classification with high sensitivity and specificity of its constituents. Furthermore, this method may be utilized by subsequent analyses such as task-specific genotyping of mice.

4.1 Action Classification

The developed automatic behavioural scoring model presented is composed of two major parts, that is one unsupervised sub-model and one supervised sub-model. The former relies on ad hoc definitions of what constitutes a behavioural model, exploiting existing pose estimation techniques. In contrast, the latter involves a self-learning decision process that uses optic flow to define what comprises a behavioural module.

The unsupervised sub-model that classifies both mouse Wall Exploration and Corner Sitting implements the recent pose estimation algorithm DeepLabCut to identify actions based on the geometrical configuration of multiple body parts (Mathis et al., 2018; Nath* et al., 2018; Insafutdinov et al., 2016). Specifically, the ears, nose, back, and tail were used to infer mouse head direction and body position allowing to quantify behaviour. This thesis supports the reliability and unexplored use cases (i.e. infer head direction, action classification) of this specific pose estimation software. To elucidate, it validates that the feature layers from the DeepLab ANN may be used to train a new model on images of mice in a box that are relatively low quality. Thus, transfer learning in pose estimation provides an array of opportunities for researchers to identify their animals without marking and moreover using this information to perform behavioural analysis. The behaviours in this model were predefined by both mouse position and head direction, which is ideal in situations where there is not a lot of labeled video data available that consists of many action labels. One caveat of this technique might be that it could be hard to control for all possible cases in complex scenarios, potentially over engineering one model that doesn't generalize well.

The supervised sub-model that classifies mouse Object Exploration implements transfer learning on the labels of an existing human action recognition ANN (Carreira and Zisserman, 2017; Szegedy et al., 2015; Yosinski et al., 2014). Specifically, the top layers of the original inflated inception network were cut and replaced by a cascade of new layers to ultimately predict the label Object Exploration. This network used a sequence of 9 frames, evenly spread over 30 frames, as input to allow optic flow and colour channels (i.e. inception modules) to carry optimized smooth flow information to final convolutional feature-and decision layers. With an AUROC higher than 98%, this model has shown to be highly sensitive and specific to the action classification category in mice exploring objects. This supports the benefit of transfer learning on a kinetics pre-trained ANN across domains, since the type of

video data of the original network (i.e. humans) and the final network (i.e. mice) in this thesis rather dissimilar. Using ANNs for automatic behavioural scoring opens up opportunities for the translational neuroscience field as it allows for identifying behavioural modules without researcher bias to define them. That is, ANNs may learn to classify a behavioural model by dynamically changing its weights for relevant spatiotemporal feature layers to emerge without the need of human intervention. Albeit this method has been shown to be great for behavioural scoring, it might not offer much insight in what constitutes a behaviour. This is because ANN are sometimes considered black-box methods because it is hard to find a consistent way of interpreting its logic in decision, albeit methods have been suggested (Benítez, Castro, and Requena, 1997; Dayhoff and DeLeo, 2001; Tzeng and Ma, 2005). Thus, although not giving insight in the structure of one behavioural module, this technique has proven to be great at markerless automatic behavioural scoring in mice.

Prior studies have either focused on classifying predefined behavioural modules or mapping undiscovered behavioural modules (De Chaumont et al., 2012; Wiltschko et al., 2015). De Chaumont et al. (2012) developed a model that could describe multiple mice on a video, where each mouse was modeled using 3 geometric primitives (i.e. head, body, tail). These inferred body parts were then used to define various (social) behavioural modules. However, this model has proved to be inconsistent in that mouse identities tend to switch. Furthermore, the model used a physics engine to find the optimal location of body parts. In contrast, the pose estimation model used in this thesis needs to such complex physics engine an can infer approximate locations of any body part that is labeled by just a few images (Mathis et al., 2018; Nath* et al., 2018; Insafutdinov et al., 2016). This multiple body part resolution may result in the ability to expand to more detailed and reliable behavioural modules, which has partly been shown in this thesis as Wall Exploration and Corner Sitting. Next to classifying behaviour body part configuration, one may use the kinetic action recognition transfer learning technique applied in this thesis to describe multiple behaviours. Although this thesis only used one output label (i.e. Object Exploration), there is not obvious reason that this shouldn't generalize to more labels as the original model could classify up to 600 categories (Carreira and Zisserman, 2017). Another study has focused on discovering meaningful behavioural modules by finding latent behaviours in mice (Wiltschko et al., 2015). This may be especially beneficial when one aims to discover new behavioural modules that describe global behavioural patterns in some task. Once a set of latent behavioural have been found, this could subsequently be used to in another efficient action classification model such as described in this thesis. Overall, the ideas in this thesis provide a novel and reliable stance on video action classification in rodents.

The high time resolution action classification model developed in this thesis opens doors for neuroscience research in that it may be paired with techniques such as neural recording and optogenetics. For example, one may pair neural recordings with the inferred head direction and aim to explore some relationship between signal and gaze type. Integrating video action classification time and recording time could thus allow to find correlates of behaviour with neural responses. Next to passive recording, one may want to study the direct effect op optogenetic stimulation on behaviour in some task. Whereas most methods summarize behaviour over an entire trial, the developed action classification method allows for more detailed descriptions of the effect on behaviour.

To summarize, this thesis has shown that one can engineer an accurate rodent action classification model using techniques borrowed from the machine learning field such as kinetic action recognition, transfer learning, and pose estimation. This model can be used to classify three behaviours (i.e. Object Exploration, Wall Exploration, Corner Sitting), but is in principle not limited to these categories. Furthermore, this model may be used to perform a detailed behavioural analysis due to its high time resolution.

4.2 Memory Processes In a Mouse Model of Autism

Autism is a cluster of behavioural abnormalities manifesting as impaired social behaviour, perseverance behaviours and stereotypies (Klinger et al., 2019). Among these abnormalities might be affected memory processes with impaired episodic memory deficits, and condition-dependent equal or superior semantic memory performance (Goddard et al., 2014; Gaigg, Bowler, and Gardiner, 2014). In this study, ehmt1$^{+/-}$ and ehmt1$^{+/+}$ mice performed the Object-Space task for all conditions (Genzel et al., 2018). Based on videos, behavioural features were extracted for all trials and were subsequently used to predict mouse genotype per condition for each trial.

The classifier was able to do a task-dependent genotyping of mice for each trial with only significant accuracies in the stable and overlapping condition of the Object-Space task (Genzel et al., 2018). Interestingly, these are the two conditions that contain object-location patterns that may be extracted differentially between ehmt1$^{+/-}$ and ehmt1$^{+/+}$ mice. Specifically, the accuracy in the genotype prediction of trials from the overlapping condition was the highest and this condition is characterized by one stable object location and one variable object location. A proxy of pattern extraction (i.e. memory) would be some behavioural feature that distinguishes between either object. Accordingly, top features used by the genotype classifiers in the overlapping condition are dominated by object distinguishing features such as *mini_DI* and *mini_n_explore_obj1*. With the ehmt1$^{+/-}$ mice having a higher discrimination index per minute and a lower total exploration time per object than the ehmt1$^{+/-}$ in the overlapping condition over trials, this suggest that ehmt1$^{+/-}$ mice have a more efficient way of expressing memory. To elucidate, the high discrimination index indicates that the moving object is explored more than the static one thus likely being perceived as a more novel location. The accompanying lower exploration times implies that this higher memory expression is achieved in less total time. Whether this memory expression is due to semantic or episodic processes remains to be determined, since genotype was predicted over all trials and no juxtaposition between train and test trials could be made with the current model. Next, top features used by the genotype classifiers for the trials in from the stable condition are dominated by exploration time related features such as *mini_explore_time* and *mini_obj1_time*. It must be noted that *min5_DI* is the top feature of one classifiers in the stable condition. For all time related features, there is either no difference between genotype or the ehmt1$^{+/-}$ mice mice have a lower exploration time than the ehmt1$^{+/+}$ mice. Notably, the DI does not differ over trials between genotype in the stable condition. That the top features in predicting genotype in the stable trials are mostly characterized by time comes as no surprise, since for most of the trials the objects are location invariant. This task-dependent expression of exploration times supports results from prior studies that this specific mouse model of autism have reduced exploratory behaviour (Balemans et al., 2010). Overall, the results show a condition/feature-dependent genotyping with more memory related features in the overlapping condition and more general behavioural features in the stable condition. This is in line with prior studies that autism is characterized by altered memory and improved pattern extraction in some contexts (Goddard et al., 2014; Gaigg, Bowler, and Gardiner, 2014; Benevento et al., 2017). This particular study suggests that autism is characterized by an improved memory expression by object-location discrimination patterns. The results are also in line with prior studies finding innate behavioural differences in individuals with autism, namely that those with autism show reduced exploration (Balemans et al., 2010). This particular study suggests that exploration is reduced in autism invariant of condition, but might be more meaningful in the context of a stable environment.

The aforementioned results show that the Object-Space task may provide a novel way to investigate memory processes in in both healthy and disease model mice (Genzel et al., 2018). For one, this study showed that the Object-Space task has meaningful conditions which are useful in prediction mouse genotype through some behavioural and memory features expressed by mice in single trials. The superior genotype prediction accuracies in the overlapping condition implies that this task captures behavioural expressions in mice. Specifically, it suggests that autism model mice show differential memory related behavioural expression compared to healthy mice. The high genotype prediction accuracies in the stable condition implies that this task captures innate task-dependent behavioural differences that

may or may not be related to memory. Genzel et al. (2018) have shown that the Object-Space task may distinguish episodic and semantic memory processes when looking at training versus test trials. The current study did not aim to validate this, yet expands the idea that the Object-Space task opens doors to investigate memory processes in a novel way.

The genotyping methods developed in this study is, of course, no match for existing methods based on biological measures (Kwok, 2000; Tsuchihashi and Dracopoli, 2002). The developed behavioural genotyping method merely shows that there are task-dependent behavioural differences whether these are due to some cognitive process or not. Additionally, due to the feature importances extraction, this method provides a way to take a fresh look at variables typically used in varying studies. For example, in the showcase study of the Object-Space task this method may suggest that the discrimination index, typically used as a proxy of memory, may be condition/time-dependent in its expression. This is particularly depicted by that the discrimination index after minute 3 was differed between genotype of mice in the overlapping condition, but despite being important none were different in the stable condition. This also touches a point on the usefulness of a classifier over humans in the behavioural genotyping in mice. Altough one features could significantly differ between genotype, this is not sufficient to accurately distinguish between them. That is, one variable can be different, yet still have many overlapping cases such that no threshold with provide an accuracy above chance. On the other hand, even features that do not show significant differences as a single variable between genotypes may be important in the decision. This is because the used classifiers are decision trees that approximate some non-linear function between all features to predict the genotype. That is, features may modulate each other. No single feature suffices, yet together they may be used in way to predict genotype that is hard for the human observer.

Future studies using behavioural categorization may want to include different features depending on their research question. In this particular case the features used are those that are standard in the study of memory in the Object-Space task. Potential other features, such as wall or corner exploration, were deliberately omitted due to time and usefulness concerns. There may be other memory related features that the author did not think of that may proof to be of importance in distinguishing between the ehmt1 genotypes. Furthermore, future studies that aim to provide a more nuanced description of differences in episodic and semantic memory in the Object-Space task would require to use more trials. To elaborate, this study could not make a distinctive analysis between training and test trials, since classifiers need hundreds of training data points and these were simply not available for the test trials.

Overall, this study has shown a task-dependent genotyping of an autism model of mice and healthy mice. Behavioural genotyping may be most efficient in an environment that is characterized by dynamic patterns that could be extracted through object-location discrimination. This is supported by the use of the object discrimination index in the overlapping condition trials for genotyping. Furthermore, behavioural genotyping may also be done in an environment that is characterized by stable patterns. This is supported by the use of overall exploration time features in the stable condition. The ehmt1$^{+/-}$ mice show improved and efficient memory expressions in the overlapping condition, suggesting it may be a model of high-functioning autism in mice.

4.3 Conclusion

Computerized analysis may provide an observer invariant approach to extract meaningful behavioural information from video data of rodents performing a task. The extracted information can be used to describe behavioural transitions and other relations between or within behavioural modules on a time scale that is only limited by the recorder. In this thesis such video analysis has shown that: 1. mouse genotype can be predicted using behaviours, and 2. an ehmt1 mouse autism model of mice express more memory related behaviours than its healthy controls in the Object-Space task condition, where there are object-location patterns to be extracted over trials. This suggests that ehmt1$^{+/-}$ mice have improved memory or pattern extraction over ehmt1$^{+/+}$ mice.

Bibliography

Baker, Monya (2011). "Animal models: inside the minds of mice and men". In: *Nature* 475.7354, p. 123.

Bala, Rajni and Dharmender Kumar (2017). "Classification Using ANN: A Review". In: *International Journal of Computational Intelligence Research ISSN* 13.7, pp. 973–1873.

Balemans, Monique CM et al. (2010). "Reduced exploration, increased anxiety, and altered social behavior: Autistic-like features of euchromatin histone methyltransferase 1 heterozygous knockout mice". In: *Behavioural brain research* 208.1, pp. 47–55.

Balemans, Monique CM et al. (2012). "Hippocampal dysfunction in the Euchromatin histone methyltransferase 1 heterozygous knockout mouse model for Kleefstra syndrome". In: *Human molecular genetics* 22.5, pp. 852–866.

Benevento, Marco et al. (2016). "Histone methylation by the Kleefstra syndrome protein EHMT1 mediates homeostatic synaptic scaling". In: *Neuron* 91.2, pp. 341–355.

Benevento, Marco et al. (2017). "Haploinsufficiency of EHMT1 improves pattern separation and increases hippocampal cell proliferation". In: *Scientific reports* 7, p. 40284.

Benítez, José Manuel, Juan Luis Castro, and Ignacio Requena (1997). "Are artificial neural networks black boxes?" In: *IEEE Transactions on neural networks* 8.5, pp. 1156–1164.

Bergstra, James and Yoshua Bengio (2012). "Random search for hyper-parameter optimization". In: *Journal of Machine Learning Research* 13.Feb, pp. 281–305.

Blanchard, D Caroline, Guy Griebel, and Robert J Blanchard (2003). "The Mouse Defense Test Battery: pharmacological and behavioral assays for anxiety and panic". In: *European journal of pharmacology* 463.1-3, pp. 97–116.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Carreira, Joao and Andrew Zisserman (2017). "Quo vadis, action recognition? a new model and the kinetics dataset". In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, pp. 4724–4733.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, pp. 785–794.

Dayhoff, Judith E and James M DeLeo (2001). "Artificial neural networks: opening the black box". In: *Cancer: Interdisciplinary International Journal of the American Cancer Society* 91.S8, pp. 1615–1635.

De Chaumont, Fabrice et al. (2012). "Computerized video analysis of social interactions in mice". In: *Nature methods* 9.4, p. 410.

Dere, Ekrem, Joseph P Huston, and Maria A De Souza Silva (2005). "Integrated memory for objects, places, and temporal order: evidence for episodic-like memory in mice". In: *Neurobiology of learning and memory* 84.3, pp. 214–221.

Ernst, Michael D et al. (2004). "Permutation methods: a basis for exact inference". In: *Statistical Science* 19.4, pp. 676–685.

Frankland, Paul W and Bruno Bontempi (2005). "The organization of recent and remote memories". In: *Nature Reviews Neuroscience* 6.2, p. 119.

Gaigg, Sebastian B, Dermot M Bowler, and John M Gardiner (2014). "Episodic but not semantic order memory difficulties in autism spectrum disorder: Evidence from the Historical Figures Task". In: *Memory* 22.6, pp. 669–678.

Genzel, Lisa et al. (2018). "The Object Space Task for mice and rats". In: *bioRxiv*, p. 198382.

Goddard, Lorna et al. (2014). "Development of autobiographical memory in children with autism spectrum disorders: Deficits, gains, and predictors of performance". In: *Development and Psychopathology* 26.1, pp. 215–228.

Hornik, Kurt (1991). "Approximation Capabilities of Muitilayer Feedforward Networks". In: *Neural Networks* 4, pp. 251–257.

Insafutdinov, Eldar et al. (2016). "Deepercut: A deeper, stronger, and faster multi-person pose estimation model". In: *European Conference on Computer Vision*. Springer, pp. 34–50.

Kay, Will et al. (2017). "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950*.

Kleefstra, Tjitske et al. (2006). "Loss-of-function mutations in euchromatin histone methyl transferase 1 (EHMT1) cause the 9q34 subtelomeric deletion syndrome". In: *The American Journal of Human Genetics* 79.2, pp. 370–377.

Klinger, Laura Grofer et al. (2019). "Autism spectrum disorder". In: *Child psychopathology*.

Kohavi, Ron et al. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Ijcai*. Vol. 14. 2. Montreal, Canada, pp. 1137–1145.

Kwok, Pui-Yan (2000). "High-throughput genotyping assay approaches". In: *Pharmacogenomics* 1.1, pp. 95–100.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, p. 436.

Long, Mingsheng et al. (2015). "Learning transferable features with deep adaptation networks". In: *arXiv preprint arXiv:1502.02791*.

Mathis, Alexander et al. (2018). "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning". In: *Nature Neuroscience*. URL: https://www.nature.com/articles/s41593-018-0209-y.

Mittal, Paritosh, Mayank Vatsa, and Richa Singh (2015). "Composite sketch recognition via deep network-a transfer learning approach". In: *Biometrics (ICB), 2015 International Conference on*. IEEE, pp. 251–256.

Moscovitch, Morris et al. (2016). "Episodic memory and beyond: the hippocampus and neocortex in transformation". In: *Annual review of psychology* 67, pp. 105–134.

Nath*, Tanmay et al. (2018). "Using DeepLabCut for 3D markerless pose estimation across species and behaviors". In: *bioRxiv*. DOI: 10.1101/476531. eprint: https://www.biorxiv.org/content/early/2018/11/24/476531.full.pdf. URL: https://www.biorxiv.org/content/early/2018/11/24/476531.

Nielsen, Michael A. (2015). "Neural Networks and Deep Learning". In: (visited on 01/19/2018).

Pandey, Dilip Kumar et al. (2008). "Depressant-like effects of parthenolide in a rodent behavioural antidepressant test battery". In: *Journal of Pharmacy and Pharmacology* 60.12, pp. 1643–1650.

Roberts, William A (2016). "Episodic memory: Rats master multiple memories". In: *Current Biology* 26.20, R920–R922.

Scarselli, Franco and Ah Chung Tsoi (1998). "Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results". In: *Neural networks* 11.1, pp. 15–37.

Schmidhuber, Jürgen (2015). "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* 61, pp. 85–117. ISSN: 18792782. DOI: 10.1016/j.neunet.2014.09.003.

Squire, Larry R (2004). "Memory systems of the brain: a brief history and current perspective". In: *Neurobiology of learning and memory* 82.3, pp. 171–177.

Szegedy, Christian et al. (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Tinbergen, Niko (1951). "The study of instinct." In:

Tsuchihashi, Z and NC Dracopoli (2002). "Progress in high throughput SNP genotyping methods". In: *The pharmacogenomics journal* 2.2, p. 103.

Tzeng, F-Y and K-L Ma (2005). "Opening the black box-data driven visualization of neural networks". In: *VIS 05. IEEE Visualization, 2005.* IEEE, pp. 383–390.

van Gerven, Marcel (2017). "Computational Foundations of Natural Intelligence". In: *Frontiers in Computational Neuroscience* 11, pp. 1–39. ISSN: 1662-5188. DOI: 10.3389/fncom.2017.00112.

Welch, William J (1990). "Construction of permutation tests". In: *Journal of the American Statistical Association* 85.411, pp. 693–698.

Wiltschko, Alexander B et al. (2015). "Mapping sub-second structure in mouse behavior". In: *Neuron* 88.6, pp. 1121–1135.

Yizhar, Ofer et al. (2011). "Neocortical excitation/inhibition balance in information processing and social dysfunction". In: *Nature* 477.7363, p. 171.

Yosinski, Jason et al. (2014). "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems*, pp. 3320–3328.

You, Quanzeng et al. (2015). "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks." In: *AAAI*, pp. 381–388.