Behavioural Processes 212 (2023) 104944

Contents lists available at ScienceDirect

Behavioural Processes

journal homepage: www.elsevier.com/locate/behavproc

Drosophila genotypes can be predicted from their exploration locomotive trajectories using supervised machine learning

Minh Nguyen [a], Gregg W. Roman [b,*], Benjamin Soibam [a,*]

[a] Department of Computer Science and Engineering Technology, University of Houston-Downtown, One Main St, Houston, TX 77002, USA
[b] Department of Biomolecular Sciences, School of Pharmacy, University of Mississippi, 415W Faser Hall, University, MS 38677-1848, USA

ARTICLE INFO

Keywords:
Drosophila
Habituation
Exploration
Machine learning

ABSTRACT

This study employs supervised machine learning algorithms to test whether locomotive features during exploratory activity in open field arenas can serve as predictors for the genotype of fruit flies. Because of the nonlinearity in locomotive trajectories, traditional statistical methods that are used to compare exploratory activity between genotypes of fruit flies may not reveal all insights. 10-minute-long trajectories of four different genotypes of fruit flies in an open-field arena environment were captured. Turn angles and step size features extracted from the trajectories were used for training supervised learning models to predict the genotype of the fruit flies. Using the first five minute locomotive trajectories, an accuracy of 83% was achieved in differentiating wild-type flies from three other mutant genotypes. Using the final 5 min and the entire ten minute duration decreased the performance indicating that the most variations between the genotypes in their exploratory activity are exhibited in the first few minutes. Feature importance analysis revealed that turn angle is a better predictor than step size in predicting fruit fly genotype. Overall, this study demonstrates that features of trajectories can be used to predict the genotype of fruit flies through supervised machine learning methods.

1. Introduction

Features of locomotor activity of animals in open field arenas have been used to demonstrate differences between different genotypes or even in the same genotype in response to stimuli over time and across different situations (Bell et al., 2009; Perals et al., 2017). For example, an analysis of the locomotive behavior of mice deficient in the somatostatin receptor 4 (sst4) gene, which mediates anti-depressant effects and is a target for drug development, revealed that ss4 influences locomotive and exploratory movement in young mice but not during normal aging (Szentes et al., 2019). Likewise, the analysis of locomotive activities of different alleles of the Drosophila clock gene showed that double-time gene is responsible for setting up the period of locomotor activity rhythms of fruit flies (Price et al., 1998). Increasingly, researchers find that locomotive behavioral analyses can reveal insights in studies investigating models of memory, anxiety, pain, sensorimotor control, etc. (Browne et al., 2017; Harris, 1943; Leal et al., 2015; Soibam et al., 2013).

This paper focusses the differences in locomotive exploratory activity between different genotypes of fruit flies during a common form of non-associative learning behavioral mechanism called habituation (Harris, 1943). Habituation, even though is a basic form of behavioral plasticity, is a complex mechanistic trait involving dynamic interactions between the animal's learning, memory system, and the environment. One of the widely studied behavioral patterns related to habituation is the decrease in the exploratory activity of many animal species during exposure to a novel open-field arena (Soibam et al., 2013). Exploratory behavior is motivated by the novelty of the arena is defined as a collection of acts and postures that allows an animal to gather information on a new environment (Liu et al., 2007; Soibam et al., 2014; Soibam, Goldfeder et al., 2012; Soibam, Mann et al., 2012). Exploration behaviors decrease as the novelty subsides (Soibam et al., 2013). To understand habituation, the exploratory behavioral patterns of different genotypes of fruit flies in open field arenas has been studied and compared using traditional statistical tests. Mutant genotypes consistently show deviations in features of locomotive trajectories of wild-type fruit flies during exploratory activity (Soibam et al., 2013). This observation provides a platform to further link these quantitative behavioral observations in habituation to genetic loci using QTL mapping studies. However, locomotive trajectories of fruit flies during habituation may have non-linear components which may not be addressed completely by some of the statistical tests. These methods cannot test whether trajectory features

* Corresponding authors.
E-mail addresses: groman@olemiss.edu (G.W. Roman), soibamb@uhd.edu (B. Soibam).

https://doi.org/10.1016/j.beproc.2023.104944
Received 9 December 2022; Received in revised form 25 August 2023; Accepted 13 September 2023
Available online 16 September 2023

can serve as predictors to discriminate different genotypes of fruit flies with different forms of exploratory behavior. Supervised machine learning methods are ideal for such kind of purpose which involves building a mathematical model that predicts the label of input data based on some features of the input data. In recent years machine learning methodologies have been employed in many behavioral studies (Berman et al., 2014; Branson et al., 2009; Dankert et al., 2009).

In this paper, supervised machine learning methods were employed to predict the genotype of fruit flies based on the features of their locomotive trajectories inside an open field arena. Prediction of genotype and species of small insects is of immense importance in many areas of monitoring and population estimation for pest control, research in entomology, and agriculture (Cardim Ferreira Lima et al., 2020; Gerovichev et al., 2021; Høye et al., 2021). Our approach applies to predicting genotypes based only on trajectory information which doesn't require high-quality images revealing small body parts of small insects (Gerovichev et al., 2021; Høye et al., 2021). Even though we use fruit flies' movement in a laboratory setting, it can provide a proof of concept that species of small insects can be predicted by relying on the movement trajectories. Besides these applications in different areas, using supervised models in the genotype prediction allow deciphering and ranking behavioral features that can distinctly discriminate the different genotypes.

2. Methods and materials

2.1. Fly stocks and husbandry

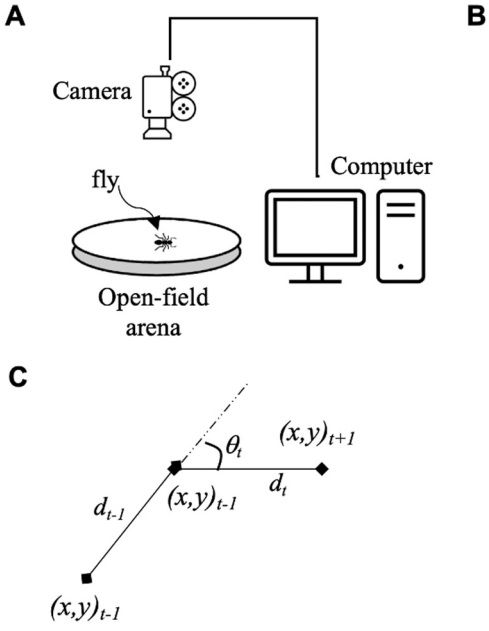
Fly stocks and husbandry details have been described in previous studies (Soibam et al., 2012; Soibam, Mann et al., 2012; Soibam et al., 2013). All stocks were raised and maintained on standard yeast-cornmeal agar food at room temperature. Flies that were raised on standard food at 25^0 C, 60% humidity, with 12 hr of light/dark. The four Drosophila genotypes used were wild-type Canton-S, and three mutants: *norpA*^{7}, *rutabaga*^{2080}, and w^{1118}. The number of experiments for each genotype is shown in Fig. 1A. The *norpA*^{7} mutant flies are defective in phospholipase Cβ, fail to perform a receptor potential, and are completely blind (Harris and Stark, 1977). The *rutabaga*^{2080} mutants are defective in a type I adenylyl cyclase and have learning defects (Lebreton and Martin, 2009). The w^{1118} mutants have mutations in the *white* gene leading to increased sensitivity to light and decreased visual acuity (Ferreiro et al., 2018). The *norpA*^{7} mutants were obtained from the Bloomington Stock Center. The *rutabaga*^{2080} and w^{1118} mutants were obtained from Ronald Davis (Scripps FL). The mutations were all crossed into a wild-type Canton-S genotype for a minimum of 6 generations.

2.2. Trajectories

As described in previous studies (Soibam et al., 2012; Soibam, Mann et al., 2012; Soibam et al., 2013), a circular open-field arena was used to collect the trajectories of fruit flies (Fig. 1A). It was made of transparent plexiglass by the University of Houston Physics Machine shop. The circular arena was 0.7 cm in height and 4.2 cm in radius. The arena's top was a lid of 15-cm Petri plates (Fisher Scientific). The aspiration of a fly into the arena was done through a 2-mm hole in the lid top (Soibam et al., 2013). Once the fly was introduced, the hole was shifted out of the active arena area to prevent the fly from escaping. The arena was illuminated by two 23 W compact fluorescent floodlights (R40, 1200 lumens, 5100 K) (Soibam et al., 2012; Soibam, Mann et al., 2012; Soibam et al., 2013). Ethovision XT v5.0 (Noldus Information Technology, Leesburg VA) was used to track and extract the (x,y) locations of the fly within the arena at a recording rate of 30 frames per second for 10 min. Each trajectory was discretized with a time unit of 1 s, and motion within this time interval was assumed to be linear. Therefore, the trajectory of a fruit fly can be represented by a sequence of (x,y) locations for 600 time steps: $\{(x,y)_1, (x,y)_2, \ldots.(x,y)_t, (x,y)_{t+1}, \ldots\ldots(x,y)_T\}$, where $T = 600$. The total number of experiments for each fly genotype is shown in Fig. 1B. We assumed the locomotion to be linear between two consecutive time points.

2.3. Calculation of step size and turn angle

The step size at time t (d_t), was the distance the fly moved between time t and $t + 1$. It was calculated as the Euclidean distance between positions of the fly at time t and $t + 1$ ($(x,y)_t$ and $(x,y)_{t+1}$) (Fig. 1C). The positions $(x,y)_{t-1}$, $(x,y)_t$, $(x,y)_{t+1}$ at three consecutive time-points t-1, t, and $t + 1$, respectively were used to compute the turn angle (θ_t) at time t using the cosine rule: $(R_{t-1,t+1})^2 = (R_{t-1,t})^2 + (R_{t,t+1})^2 + 2(R_{t-1,t})(R_{t,t+1})cos$



B

Genotype	Number of Experiments	Genotype Features
Canton-S	275	Wild type
norpA^7	120	Phospholipase CB defect Blind
rutabaga^{2080}	62	Type I adenyl cyclase and pleiotropic learning defects
w^{1118}	72	Poor Visual contrast and cannot perform optomotor tasks

Fig. 1. : Fruit fly genotypes used in this study. (A) An illustration showing how the trajectories of fruit flies inside a circular open-field arena were collected. (B) Four genotypes of fruit flies and their characteristics. (C) Calculation of turn angle (θ_t) and step size (d_t) at a specific time point t in a trajectory is shown.

(180^0 - θ_t), where ($R_{t,t'}$) is the Euclidean distance between positions $(x,y)_t$ and $(x,y)_{t'}$ (Fig. 1C).

2.4. Features

In supervised machine learning, given a set of N training examples of the form $\{(z_1, g_1), \ldots, (z_N, g_N)\}$ such that z_i is the feature vector (of some length n) of the i^{th} example and g_i is its label (i.e class label), a learning algorithm finds a function that maps/predict the label g_i based on the feature vector z_i. In our context, N is the number of experiments (flies), g_i is the genotype of the fly used in the i^{th} experiment, z_i contains turn angles and step sizes of the i^{th} fly at different time points. Since the duration of one experiment was 10 min (600-time points) long, the feature vector consisted of 598 turn angles and 598 step sizes. This was because the calculation of turn angle required three-time points (Fig. 1C).

2.5. Models and training

The goal was to test whether supervised machine learning models can be used to accurately differentiate wild-type Canton-S flies from "non-wild-type" flies based on features from the trajectories such as turn angles and step sizes (Fig. 2A). No other anatomical features such as body or wing size were used. Since, there was a lesser number of experiments for mutant flies *norpA*^{7}, *rutabaga*, and w^{1118} flies, we simply posed a binary classification problem of predicting the genotype of fruit flies (class label = 1 or 0) by using turn angles and step sizes as features. Class label 1 represented Canton-S flies, while the three other types of flies were assigned a single class label 0 (Fig. 2B). This means there were 275 and 254 experiments belonging to class labels 1 and 0, respectively. The total 529 experiments were split into training (80%) and testing sets (20%) (Fig. 2B). The split was done in such a way that a roughly equal number of experiments from class labels 1 and 0 ended up in the training set. Feature scaling in the training set was performed so that the mean and standard deviation of each feature were adjusted to 0 and 1, respectively. The parameters obtained from the scaling training set were used to scale the features in the testing set so that the training process of the model is independent of the testing set. Besides considering the features from the entire 10-minute duration, we also considered five different scenarios where the turn angles and step sizes were in the first 2.5, and 5 min, the last 2.5 and 5 min, and the entire 10 min (Fig. 2B).

A 5-fold cross-validation sampling technique was used on the training set to train five different supervised machine learning models (Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting Classifier, and Explainable Boosting Classifier). Optimal values of parameters in these models were obtained by performing hyper-parameter optimization using "grid-search" over parameter space. The parameter space for the models is provided in Table 1. For Explainable Boosting Classifier, the default parameters' values were used. Accuracy was used as the metric to decide the optimal model. The accuracies of the models on the testing set were reported for comparison. The training was done using Python and the scikit-learn package (Pedregosa et al., 2012). For the Explainable Boosting classifier, the "interpret" Python library was used.

Different supervised models used in this study are well-known methods and were selected to compare their performances (Durugkar

Table 1
Parameter space in models explored by cross-validation method. The first column represents the model. The second column represents the parameter space that was explored. Names of the parameters (based on scikit-learn python package) and their values are provided.

Model	Parameter space
Logistic Regression	• Penalty: l1, l2 • Random_state: 0, 42 • Solver: newton-cg, lbfgs, sag, saga • max_iter: 100,200,300,1000 • C: 0.1, 1, 10, 100
Support Vector Machine	• Gamma: scale, auto • Kernel: linear, rbf
Random Forest	• Criterion: gini, entropy • Max_features: auto, sqrt, log2 • Max_samples: 0.25, 0.5, 0.75, 1.0
Gradient Boosting Classifier	• loss: deviance, exponential • n_estimators: 500, 1000 • max_depth: 3, 5

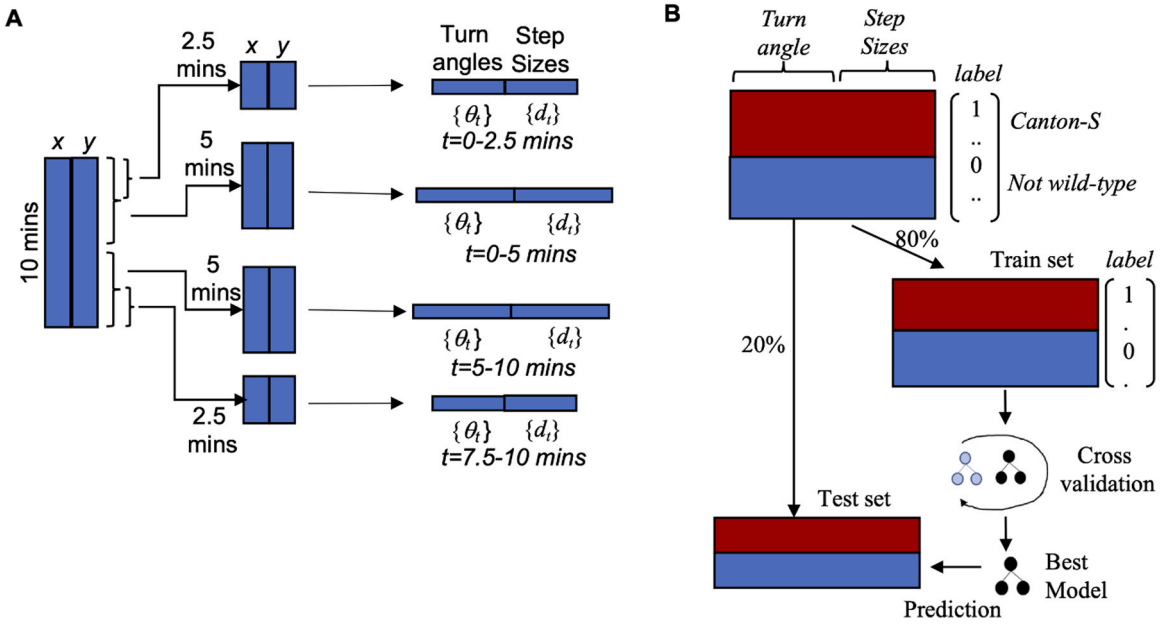


Fig. 2. : Features and supervised machine learning model training. (A) An illustration showing the extraction of turn angle and step size features from a trajectory is shown. Besides the 10-minute duration, different smaller sections of the entire 10-minute duration were also considered. (B) Training, validation, and evaluation of a model for fruit fly genotype prediction. For the binary classification problem, fruit flies were grouped into two classes: "wild type" or class label "1" and "non-wild type" or class label "0".

et al., 2022; Ruppert, 2004). A simple model like logistic regression performs well when the data points from different class labels are linearly separated (Durugkar et al., 2022; Ruppert, 2004). In our context, this can be used to test if the trajectories feature linearly separate the different genotypes. Support vector machines and ensemble models such as Random Forest and Gradient Boosting have been shown to perform well on data sets with a high number of features (Durugkar et al., 2022; Ruppert, 2004). Finally, Explainable Boosting was also chosen because it allows interpretation and ranking of the features used in prediction (Lou et al., 2012).

2.6. Feature importance

The importance of each of trajectory features (turn angles and step sizes) was calculated using the Explainable Boosting Machine (Hastie and Tibshirani, 1987; Lou et al., 2012, 2013). Explainable Boosting Machine (EBM) is a tree-based, cyclic gradient boosting Generalized Additive Model with automatic interaction detection (Hastie and Tibshirani, 1987; Lou et al., 2012, 2013) and the model can also provide the contribution of each feature to a final prediction (Hastie and Tibshirani, 1987; Lou et al., 2012, 2013).

2.7. Class imbalance and multi-genotype classification

For the multi-genotype classification problem, the goal was to predict the genotype of the fruit fly as one of the four options: Canton-S, *norpA*^{7}, *w*^{1118}, or *rutabaga*^{2080}. Since the number of experiments differed across these four genotypes, the training set was not "class balanced". To prevent the supervised learning model to be biased towards the majority class (Canton-S), we used the sampling strategy called SMOTE (Synthetic Minority Over-sampling Technique) to balance the training set (Sugimura et al., 2008). Python package "imbalanced-learn" was used for this purpose (Lemaitre et al., 2016). We resample all three classes (*norpA*^{7}, *w*^{1118}, or *rutabaga*^{2080}) except the majority class (Canton-S). The testing set was kept imbalanced.

3. Results

3.1. Turn angle and step size are dependent on fly genotype but also on time

To investigate whether features of trajectories (step sizes and turn angles) of fruit flies inside an open field arena are dependent on genotype and time, we divided the 10-minute duration trajectory into four different time sections: 0–2.5 min, 2.5–5 min, 5–7.5 min, and 7.5–10 min, and explore whether the turn angles and step sizes were significantly affected by the time sections in the trajectories.

Between any two different time sections, the step sizes for each genotype were significantly different (Fig. 3A, Table S1, Kruskal-Wallis: p-value < 0.001). This means the step size decreased significantly as time progressed for each genotype. Canton-S had the largest decrease (0.58 cm) in step size from 0.73 cm during the first 2.5 min to 0.15 cm during the final 2.5 min (Fig. 3A). In contrast, mutant genotypes had a smaller decrease in the step size: 0.30 cm, 0.31 cm, 0.26 cm for *norpA*^{7}, *w*^{1118}, and *rutabaga*^{2080}, respectively (Fig. 3A). Next, we also confirmed that during a specific time section of the 10-minute duration, the step sizes between any two genotypes of the fly were significantly different (Fig. 3A, Table S2, Kruskal-Wallis test, p-value < 0.001). In each time

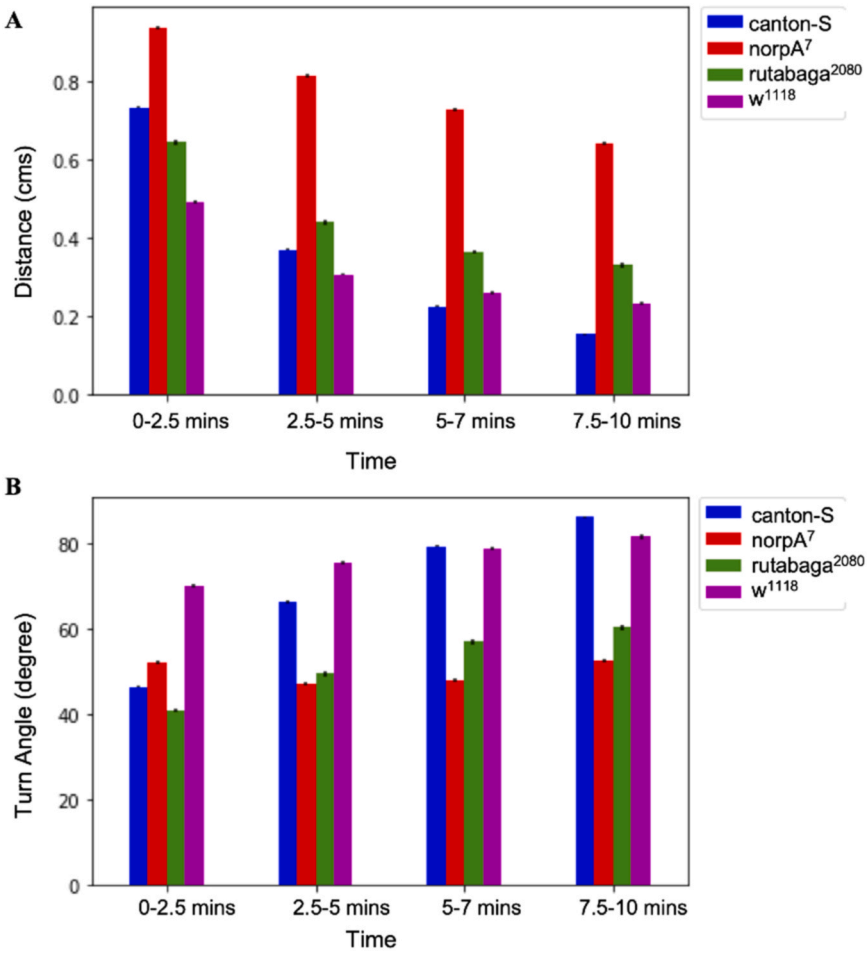


Fig. 3. : Turn angle and step size depend on time and genotype. Mean step size (cms) and turn angle (degrees) during different time sections of the 10-minute experiment are shown as bar plots in (A) and (B), respectively. Error bars are standard errors of the mean. Genotypes are colored-coded. The x-axis indicates the four different time sections, and the y-axis indicates step size (panel A) or turn angle (panel B). Statistical analyses testing the effect of genotype and time on the turn angle and step size are shown in Tables S1, S2, S3, and S4.

section, *norpA*^{7} mutant flies showed the highest step sizes (Fig. 3A).

Except for three comparisons (*norpA*^{7}*:* 2.5–5 mins vs 5–7.5 mins, *rutabaga*^{2080}: 0–2.5 mins vs 2.5–5 mins and 5–7.5 mins vs 7.5–10 mins), the turn angles for each genotype were significantly different between any two different time sections (Fig. 3B, Table S3, Kruskal-Wallis: p-value < 0.001). In general, the turn angle increased as time increased for each genotype (Fig. 3B). However, the increase in turn angle was more significant in Canton-S flies compared to the mutant flies. For example, the Canton-S flies exhibited an average turn angle of 46^{o}, 66 ^{o}, 79 ^{o}, and 85^{o} during the 0–2.5 min, 2.5–5 min, 5–7.5 min, and 7.5–10 min sections, respectively (Fig. 3B). The *rutabaga*^{2080} mutant flies showed averaged turn angles of 41^{o}, 49 ^{o}, 57 ^{o}, and 60^{o} during the four different time sections (Fig. 3B). Except for Canton-S and w^{1118} at 5–7 min section, any two genotypes exhibited different turn angles during each time section (Fig. 3B, Table S4, Kruskal-Wallis test, p-value < 0.001). During the 0–2.5 min and 2.5–5 min sections, w^{1118} had the largest turn angles (Fig. 3B). However, during the 5–7.5 min and 7.5–10 min sections, Canton-S had the largest turn angles because Canton-S showed the largest increase in their turn angles with time (Fig. 3B). Overall, these results indicate that not only turn angle and step size are dependent on fly genotype but also on time.

3.2. Turn angle and step size predict the genotype of fruit fly

Since turn angles and step sizes were different across the genotypes of the fruit flies, we aimed to test whether the genotype of a fruit fly can be predicted from turn angles and step sizes associated with their locomotive trajectories. Since there was an unequal number of experiments across the four different genotypes, a binary classification problem was posed where the genotype or class label of the fly (class label 1: "wild type/Canton-S" or class label 0: "not wild type/Canton-S") can be predicted using turn angles and step sizes from the trajectories. There were 275 and 254 (from three mutant genotypes) experiments belonging to class labels 1 and 0, respectively. This classification was done to have a balanced number of samples across the two class labels.

To differentiate Canton-S flies from three mutant genotypes, we used five different supervised machine learning models (Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, and Explainable Boosting Classifier) with turn angles and step sizes at different time points during the 10-minute duration. Besides utilizing these measures from the entire 10-min duration, we also considered shorter time intervals: first 2.5 and 5 min, final 2.5, and 5 min (Fig. 2A). The training, validation, and testing of these models were performed using a 5-fold cross-validation method (Fig. 2B). When the first 2.5 min of the trajectories were considered, the Gradient Boosting classifier yielded the highest accuracy of 78%, while logistic regression obtained only 58% accuracy (Fig. 4). The accuracy of other models ranged between 70% and 75% (Fig. 4). On the other hand when the first 5 min of the trajectories were considered, Explainable Boosting Classifier achieved the highest accuracy of 83% followed by Gradient Boosting with 80% (Fig. 4). The Logistic Regression model had the lowest accuracy of 60% (Fig. 4). Random Forest and Support Vector Machines achieved 75% and 78% accuracies, respectively (Fig. 4). The accuracies of all models dropped when the final 2.5 and 5 min were considered. In these cases, the best accuracies achieved were 76% (Gradient Boosting) and 72% (Explainable Boosting Classifier), respectively (Fig. 4). The reduction in accuracies indicates that the variation in turn angle and step sizes across the genotypes in the first few minutes was more significant than the final few minutes in the trajectories. Training the models with the entire 10-minute duration reached 82% accuracy with the Explainable Gradient Boosting Classifier (Fig. 4). Logistic regression was the worst model, which suggests that there is a non-linear relationship between turn angles/step sizes and the genotype of fruit flies (Fig. 4). To check for overfitting or underfitting, we focused on the best model (explainable Boosting Machine on 5 min data). It had comparable accuracies in training and testing data of 86% and 83%, respectively indicating that the trained model was able to generalize to the testing set. Overall, our results indicate that step sizes and turn angles are sufficient to differentiate wild-type Canton-S flies from other mutant genotype fruit flies.

3.3. Turn angle is a better predictor than step size for the genotype of fruit fly

The Explainable Boosting Machine yielded the highest accuracies among the models. It is also a highly interpretable model that yields each feature's importance score based on the training set. Therefore, we used this model to evaluate the importance of each feature in differentiating wild-type flies from other mutant genotypes. When the first 5 min were considered, turn angles had the highest importance scores during the first minute (Fig. 5). The importance scores of turn angles decreased as time increased (Fig. 5). Unlike turn angle, there was no clear observable pattern in the importance scores of the step size features (Fig. 5). In general, the importance score of step size remained roughly the same throughout the 5-minute duration (Fig. 5). Similar observations were made when the first 2.5-minute duration was considered (Fig. 5). These results indicate that turn angles of fruit flies are better predictors for genotype compared to step sizes. Interestingly, when the final 2.5 or 5 min were considered, the importance scores of turn angles and step sizes did not follow an increasing or decreasing trend throughout the 2.5- or 5-minute duration (Fig. 5). This was most likely because the features during the final minutes are not sufficient to distinguish the genotypes.

Since the turn angles were better predictors compared to the step sizes, we further checked whether turn angles and step sizes displayed any correlation and whether using turn angles only improved model performance. We took the first 5 min of the trajectories and computed the Pearson correlation coefficient between turn angles and step sizes for

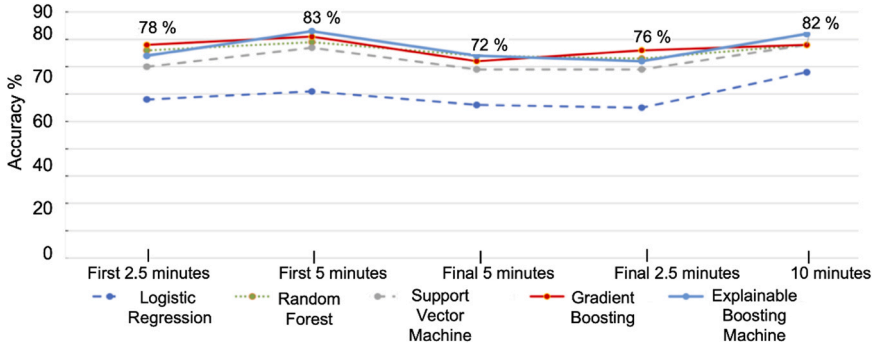


Fig. 4. : Accuracies of different supervised machine learning models. Accuracies of five different models are indicated and differentiated by different line types. The y-axis indicates the accuracies of the models in differentiating Canton-S flies from three mutant flies. Five different cases where different portions of the 10-minute duration are represented along the x-axis. For each case, the highest accuracy % is indicated.

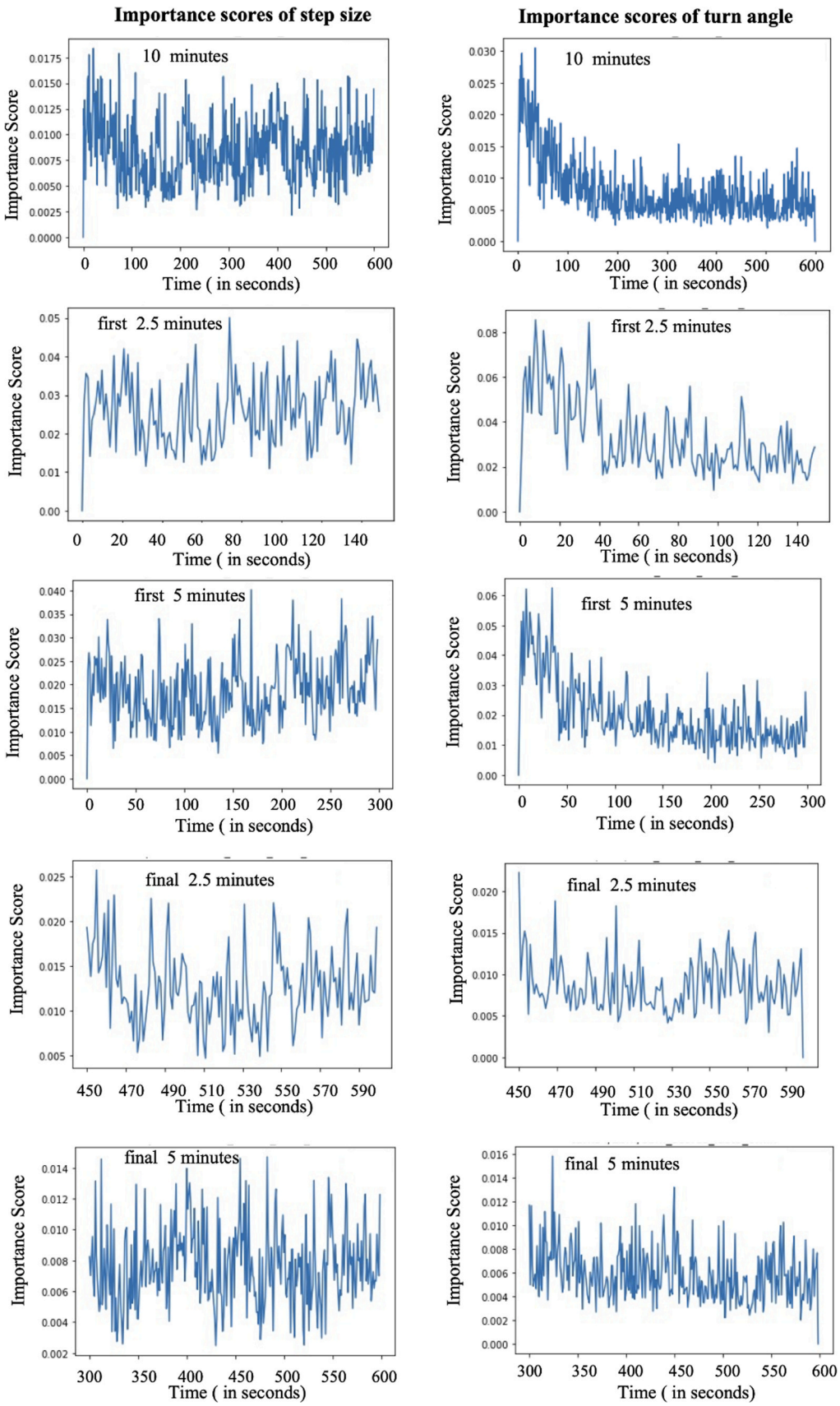


Fig. 5. : Importance score of turn angle and step size in predicting fruit fly genotype. The x-axis indicates the time points (in seconds), and the y-axis indicates importance scores obtained using Explainable Boosting Classifier.

each fly. The averaged Pearson correlation coefficient was − 0.14 indicating that there may be a small negative correlation between the turn angles and step sizes. This may be because flies need to slow down while executing larger turn angles. To test whether using only turn angles improve model performance, we trained the Explainable Boosting Classifier model using only the turn angles in the first five minutes. This yielded an accuracy of 80% in binary genotype classification which was slightly less than the 83% accuracy achieved by using both turn angles and step sizes features. Interestingly, only using step sizes achieved a much-reduced accuracy of 72%. This confirms that turn angles are better predictors than step sizes. However, using both turn angles and step sizes increases model performance compared to using only turn angles.

3.4. Prediction accuracy decreased in the multi-genotype classification setting

Instead of a binary classification problem of predicting "wild type/Canton-S" from "non-wild type" flies, we posed a multi-genotype classification problem of predicting the genotype of the fly from one of the four possible genotypes. Because of the imbalance in the number of experiments across the four genotypes, the training set was balanced using a resampling technique called SMOTE (Sugimura et al., 2008). The testing set was kept unbalanced. Because of its superior performance, Explainable Boosting Machine was used for training, validation, and testing. The best performance occurred when the first 5 min were considered achieving 80%, 67%, 60%, and 56% accuracies in predicting the four genotypes Canton-S, *norpA*^{7}, *w*^{1118}, or *rutabaga*^{2080}, respectively (Table 2). The accuracy was the highest for Canton-S and lowest for the *rutabaga*^{2080} mutant (Table 2). This is most likely because *rutabaga*^{2080} had the least number of experiments. The low accuracy for other mutant genotypes is most likely because of the lesser number of experiments in the training set.

In summary, the performance of the supervised models was dependent on the time sections of the trajectories that were used in the model. Using the first 5-minute locomotive trajectories, we achieved the best accuracy of 83% in differentiating wild-type flies from three other mutant genotypes. This means that different genotypes of fruit flies exhibit the most variations in the first few minutes of their exploratory activity. Accuracy was decreased when the prediction was performed to identify each genotype in a multi-genotype prediction supervised learning problem. Feature importance analysis revealed that turn angle was a better predictor than step size in predicting fruit fly genotype. Overall, this study shows that features of trajectories can be used to predict the genotype of fruit flies.

4. Discussion

In this paper, we explored the possibility of detecting the genotype of small insects solely from trajectories by using turn angles and step sizes of fruit flies executed inside a circular open-field arena. By testing five different supervised machine-learning models, we were able to achieve an accuracy of 83% in differing wild-type flies from three other mutant genotypes. These data show that genotypes can be predicted entirely from locomotive trajectories without other anatomical information about the flies. The best accuracy was achieved when the first 5 min of locomotion inside the arena were considered. Accuracy was lower when the final 5 min were considered. This result indicates that the first few minutes showed the maximum variation in the locomotive trajectories across different genotypes. The varying locomotive trajectories exhibited by different genotypes may be closely related to the phenomena of habituation in a novel environment. Habituation is a common form of non-associative learning in which an organism gradually decreases its response to repeated stimuli (Harris, 1943; Soibam et al., 2013). In general, initial locomotive trajectories of animals show directional persistence (small turn angles) and large step sizes, which are motivated by the novelty of an environment (Soibam et al., 2013). Previous theoretical studies have shown that there is a strong relationship between movement patterns and the efficacy of animal search strategies in a novel environment (Bartumeus et al., 2008; Bartumeus and Levin, 2008; Viswanathan et al., 1999). As an animal learns the environment through repeated exposure, the directional persistence and movement decrease (Liu et al., 2007; Soibam, Mann et al., 2012; Soibam et al., 2013). However, the locomotive response can vary across different genotypes of fruit flies depending on their learning acuity. *NorpA*^{7} lacks visual learning ability and hence may not be able to learn the environment even with continued exposure (Soibam et al., 2013). Therefore, these mutant flies failed to habituate and showed the least amount of change in step sizes and turn angles during the 10-minute duration.

Feature analysis with explainable gradient boosting revealed that turn angle was more important than step size. This finding resonates with studies reporting that turn angle is an essential feature of animal locomotive behavior (Bartumeus et al., 2008; Bartumeus and Levin, 2008). Animals tend to execute their movements in a new environment to maximize the net gain of resources, usually energy (Bartumeus et al., 2008; Bartumeus and Levin, 2008; Viswanathan et al., 1999). It has been shown that the frequency and extent of turns are pivotal in employing a search strategy (Janson and Bitetti, 1997; Vasquez, 2002). Considering different step size with random turn angles are not enough to accurately describe realistic animal movements (Wilson et al., 2013). The timing and amount of turn angles are critical, and it is most likely dependent on different factors including the genotype of the animal.

This paper shows that turn angles and step sizes extracted from trajectories can be used as predictors for genotype discrimination in supervised machine-learning methods. This idea may have applications in areas of pest management. Pest management involves the application of a proper number of insecticides at accurate locations. This requires proper monitoring and estimation of pest count. Current methods rely on computer vision and high-quality images of pests for monitoring purposes. In a study to identify fruit fly *Drosophila suzukii* from static images, optimal results were observed only when the image quality was sufficient, i.e., when the black spots on the wings of flies were visible to the naked eye (Roosjen et al., 2020). Current off-the-shelf camera systems are not capable of collecting images of high enough quality for the detection of objects as small as some target insects like fruit flies. Our method relies only on (x,y) positions of the fruit flies and doesn't need high spatial resolution images that reveal the body parts of the fruit flies. We used trajectories of fruit flies in an open-field arena in a controlled environment. In real scenarios, the environmental factors are not consistent. Factors such as wind, sunlight, and temperature may vary and can affect the trajectories of insects. These variables should be considered while developing a predictive model.

In our study, we primarily focused on binary classification where the wild-type Canton-S flies were distinguished from three other mutants. This was done because of a lack of experiments for the three genotypes. However, a practical model needs to perform multi-class prediction. We attempted to solve the lack of experiments for the other three genotypes by synthesizing new samples using a technique called SMOTE (Sugimura et al., 2008). Unfortunately, this resulted in mediocre accuracies in predicting the mutant genotypes. This means the synthesized samples did not accurately represent the trajectories of the mutant genotypes. In our approach, we only used the simplest features of animal trajectories:

Table 2
Accuracy in multi-genotype classification. The table describes the accuracy achieved by Explainable Boosting Machine in predicting each genotype.

Time section	*Canton-S*	*norpA*^{7}	*w*^{1118}	*rutabaga*^{2080}
First 2.5 min	84%	56%	56%	56%
First 5 min	80%	67%	60%	56%
Final 2.5 min	73%	61%	36%	56%
Final 5 min	64%	67%	44%	56%
Entire 10 min	77%	72%	48%	56%

turn angles and step sizes. In general, traditional machine learning models used in this study assumed that the turn angles and step sizes at different time points were independent of each other, which is not the case in animal trajectories. There is some dependency between turn angles and step sizes across a certain period. Deep learning models such as convolution and recurrent neural networks can leverage this fact and may be able to improve the detection of different genotypes. Deep learning models can also learn different locomotive patterns that differentiate different genotypes. However, a large training set is required for such models and will require acquiring a significant number of trajectories. These issues will be addressed in future studies by performing more experiments with a wider variety of genotypes of fruit flies.

5. Conclusions

The study demonstrates that by relying only on the features of trajectories of fruit flies, supervised machine learning methods can discriminate genotypes of fruit flies. The features of the trajectories that were used were step sizes and turn angles. The performance of the supervised models was dependent on the time section of the trajectory that was used in the models. Utilizing the first 5 min of locomotion yields the best prediction accuracy compared to the entire ten minutes or the final 5 min of the trajectories. This indicates that the variations in the exploratory activity across genotypes are most prevalent in the first few minutes of introduction to the novel arena. Turn angles served as better predictors in the supervised methods compared to the step sizes. We also found that in a multi-genotype supervised classification problem, more data acquisition to train the models may improve model performance.

Data Availability

Data will be made available on request.

Acknowledgments

We thankfully acknowledge Claire Manson-Bishop, Milena Lobaina, Rachel Gamblin, and Yuan Yuan Kang for technical assistance. We are grateful to Randy Clark and Jose Baez-Franceschi for their assistance in manufacturing the arenas. This work was supported by the National Institute of Mental Health [grant number R15 MH121859] and the National Science Foundation [grant number NSF 2135305].

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.beproc.2023.104944.

References

Bartumeus, F., Levin, S.A., 2008. Fractal reorientation clocks: Linking animal behavior to statistical patterns of search. Proc. Natl. Acad. Sci. USA 105 (49), 19072–19077. https://doi.org/10.1073/pnas.0801926105.

Bartumeus, F., Catalan, J., Viswanathan, G.M., Raposo, E.P., da Luz, M.G.E., 2008. The influence of turning angles on the success of non-oriented animal searches. J. Theor. Biol. 252 (1), 43–55. https://doi.org/10.1016/j.jtbi.2008.01.009.

Bell, A.M., Hankison, S.J., Laskowski, K.L., 2009. The repeatability of behaviour: a meta-analysis. Anim. Behav. 77 (4), 771–783. https://doi.org/10.1016/j.anbehav.2008.12.022.

Berman, G.J., Choi, D.M., Bialek, W., Shaevitz, J.W., 2014. Mapping the stereotyped behaviour of freely moving fruit flies. J. R. Soc., Interface 11 (99). https://doi.org/10.1098/rsif.2014.0672.

Branson, K., Robie, A.A., Bender, J., Perona, P., Dickinson, M.H., 2009. High-throughput ethomics in large groups of Drosophila. Nat. Methods 6 (6), 451–457. https://doi.org/10.1038/nmeth.1328.

Browne, L.E., Latremoliere, A., Lehnert, B.P., Grantham, A., Ward, C., Alexandre, C., Costigan, M., Michoud, F., Roberson, D.P., Ginty, D.D., Woolf, C.J., 2017. Time-resolved fast mammalian behavior reveals the complexity of protective pain responses. Cell Rep. 20 (1), 89–98. https://doi.org/10.1016/j.celrep.2017.06.024.

Cardim Ferreira Lima, M., Damascena de Almeida Leandro, M.E., Valero, C., Pereira Coronel, L.C., Gonçalves Bazzo, C.O., 2020. Automatic detection and monitoring of insect pests—a review. Agriculture 10 (5), 161. https://doi.org/10.3390/agriculture10050161.

Dankert, H., Wang, L., Hoopfer, E.D., Anderson, D.J., Perona, P., 2009. Automated monitoring and analysis of social behavior in Drosophila. Nat. Methods 6 (4), 297–303. https://doi.org/10.1038/nmeth.1310.

Durugkar, S.R., Raja, R., Nagwanshi, K.K., Kumar, S., 2022. Introduction to Data Mining. In Data Mining and Machine Learning Applications. Wiley, pp. 1–19. https://doi.org/10.1002/9781119792529.ch1.

Ferreiro, M.J., Pérez, C., Marchesano, M., Ruiz, S., Caputi, A., Aguilera, P., Barrio, R., Cantera, R., 2018. Drosophila melanogaster white mutant w1118 undergo retinal degeneration. Front. Neurosci. 11 https://doi.org/10.3389/fnins.2017.00732.

Gerovichev, A., Sadeh, A., Winter, V., Bar-Massada, A., Keasar, T., Keasar, C., 2021. High throughput data acquisition and deep learning for insect ecoinformatics. Front. Ecol. Evol. 9 https://doi.org/10.3389/fevo.2021.600931.

Harris, J.D., 1943. Habituatory response decrement in the intact organism. Psychol. Bull. 40 (6), 385–422. https://doi.org/10.1037/h0053918.

Harris, W.A., Stark, W.S., 1977. Hereditary retinal degeneration in Drosophila melanogaster. A mutant defect associated with the phototransduction process. J. Gen. Physiol. 69 (3), 261–291. https://doi.org/10.1085/jgp.69.3.261.

Hastie, T., Tibshirani, R., 1987. Generalized additive models: some applications. J. Am. Stat. Assoc. 82 (398), 371–386. https://doi.org/10.1080/01621459.1987.10478440.

Høye, T.T., Ärje, J., Bjerge, K., Hansen, O.L.P., Iosifidis, A., Leese, F., Mann, H.M.R., Meissner, K., Melvad, C., Raitoharju, J., 2021. Deep learning and computer vision will transform entomology. Proc. Natl. Acad. Sci. USA 118 (2). https://doi.org/10.1073/pnas.2002545117.

Janson, C.H., Bitetti, M.S. Di, 1997. Experimental analysis of food detection in capuchin monkeys: effects of distance, travel speed, and resource size. Behav. Ecol. Sociobiol. 41 (1), 17–24. https://doi.org/10.1007/s002650050359.

Leal, G., Afonso, P.M., Salazar, I.L., Duarte, C.B., 2015. Regulation of hippocampal synaptic plasticity by BDNF. Brain Res. 1621, 82–101. https://doi.org/10.1016/j.brainres.2014.10.019.

Lebreton, S., Martin, J.-R., 2009. Mutations affecting the camp transduction pathway disrupt the centrophobism behavior. J. Neurogenet. 23 (1–2), 225–234. https://doi.org/10.1080/01677060802509160.

Lemaitre, G., Nogueira, F., Aridas, C.K., 2016. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. ⟨http://arxiv.org/abs/1609.06570⟩.

Liu, L., Davis, R.L., Roman, G., 2007. Exploratory activity in Drosophila requires the kurtz nonvisual arrestin. Genetics 175 (3), 1197–1212. https://doi.org/10.1534/genetics.106.068411.

Lou, Y., Caruana, R., Gehrke, J., 2012. Intelligible models for classification and regression. Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 150–158. https://doi.org/10.1145/2339530.2339556.

Lou, Y., Caruana, R., Gehrke, J., Hooker, G., 2013. Accurate intelligible models with pairwise interactions. Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 623–631. https://doi.org/10.1145/2487575.2487579.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2012. Scikit-learn: machine learning in python. J. Mach. Learn. Res. ⟨http://arxiv.org/abs/1201.0490⟩.

Perals, D., Griffin, A.S., Bartomeus, I., Sol, D., 2017. Revisiting the open-field test: what does it really tell us about animal personality? Anim. Behav. 123, 69–79. https://doi.org/10.1016/j.anbehav.2016.10.006.

Price, J.L., Blau, J., Rothenfluh, A., Abodeely, M., Kloss, B., Young, M.W., 1998. double-time is a novel drosophila clock gene that regulates PERIOD protein accumulation. Cell 94 (1), 83–95. https://doi.org/10.1016/S0092-8674(00)81224-6.

Roosjen, P.P.J., Kellenberger, B., Kooistra, L., Green, D.R., Fahrentrapp, J., 2020. Deep learning for automated detection of Drosophila suzukii: potential for UAV -based monitoring. Pest Manag. Sci. 76 (9), 2994–3002. https://doi.org/10.1002/ps.5845.

Ruppert, D., 2004. The elements of statistical learning: data mining, inference, and prediction, 567–567 J. Am. Stat. Assoc. 99 (466). https://doi.org/10.1198/jasa.2004.s339.

Soibam, B., Mann, M., Liu, L., Tran, J., Lobaina, M., Kang, Y.Y., Gunaratne, G.H., Pletcher, S., Roman, G., 2012. Open-field arena boundary is a primary object of exploration for Drosophila. Brain Behav. 2 (2), 97–108. https://doi.org/10.1002/brb3.36.

Soibam, B., Goldfeder, R.L., Manson-Bishop, C., Gamblin, R., Pletcher, S.D., Shah, S., Gunaratne, G.H., Roman, G.W., 2012. Modeling Drosophila positional preferences in open field arenas with directional persistence and wall attraction. PLoS ONE 7 (10).

Soibam, B., Shah, S., Gunaratne, G.H., Roman, G.W., 2013. Modeling novelty habituation during exploratory activity in Drosophila. Behav. Process. 97, 63–75. https://doi.org/10.1016/j.beproc.2013.04.005.

Soibam, B., Chen, L., Roman, G.W., Gunaratne, G.H., 2014. Exploratory activity and habituation of Drosophila in confined domains. Eur. Phys. J. Spec. Top. 223 (9), 1787–1803. https://doi.org/10.1140/epjst/e2014-02226-7.

Sugimura, T., Arnold, E., English, S., Moore, J., 2008. Chronic suprapubic catheterization in the management of patients with spinal cord injuries: analysis of upper and lower urinary tract complications. BJU Int. 101 (11), 1396–1400. https://doi.org/10.1111/j.1464-410X.2007.07404.x.

Szentes, N., Tékus, V., Mohos, V., Borbély, É., Helyes, Z., 2019. Exploratory and locomotor activity, learning and memory functions in somatostatin receptor subtype 4 gene-deficient mice in relation to aging and sex. GeroScience 41 (5), 631–641. https://doi.org/10.1007/s11357-019-00059-1.

Vasquez, R.A., 2002. The influence of habitat on travel speed, intermittent locomotion, and vigilance in a diurnal rodent. Behav. Ecol. 13 (2), 182–187. https://doi.org/10.1093/beheco/13.2.182.

Viswanathan, G.M., Buldyrev, S.V., Havlin, S., da Luz, M.G.E., Raposo, E.P., Stanley, H. E., 1999. Optimizing the success of random searches. Nature 401 (6756), 911–914. https://doi.org/10.1038/44831.

Wilson, R.P., Griffiths, I.W., Legg, P.A., Friswell, M.I., Bidder, O.R., Halsey, L.G., Lambertucci, S.A., Shepard, E.L.C., 2013. Turn costs change the value of animal search paths. Ecol. Lett. https://doi.org/10.1111/ele.12149.