Predicting genotype of fruit flies from locomotive trajectories using supervised Machine Learning

Minh Nguyen [1], Gregg Roman[2], Benjamin Soibam[1]

[1]Computer Science and Engineering Technology, University of Houston-Downtown, [2] Department of Biomolecular Sciences, School of Pharmacy, University of Mississippi

Introduction

Many studies in neuroscience and translational analysis of disease models use quantitative analysis of locomotive behavior from different animal genotypes to decipher and comprehend neurological mechanisms, the efficiency of new therapeutic techniques, or effect of a gene mutation. However, there has been no study that attempts to predict the genotype of an animal entirely from their locomotive trajectories. Such studies have tremendous application in area of integrated pest management (IPM), which requires monitoring and estimation of pest populations. Automated monitoring approach based on computer vision rely on high quality images of these insects obtained from an optimized imagery system, which may not be optimal for small insects. In this paper, we explored an alternative but effective approach that relies only on locomotive trajectories. Such trajectories can be collected by a simple imaging system and doesn't require high quality images.

Fruit flies' trajectories

We used the trajectories of different genotypes of fruit flies in an open-field arena environment in a laboratory setting. Each trajectory consisted of a sequence of (x,y) locations during a 10-minute inside a circular open-field arena of radius 4.2 cm captured using a simple camera. Therefore, 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 $T = 600$.

We simply posed a binary classification problem of predicting the genotype of fruit fly (class label = 1 or 0) based on their turn angles and step sizes. Class label 1 represented canton-S flies, while the three other types of flies were assigned a single class label 0. Hence, the goal was to implement a model to accurately differentiate wild-type Canton-S flies from "non-wild-type" flies. This means there were 275 and 254 experiments belonging to class labels 1 and 0, respectively.

Table: List of fruit fly genotypes used in this study is shown.

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

Methods

Trajectory features: Step Size and Turn Angle

The step size at time t (d_t), was the distance the fly traveled 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 $(x,y)_{t+1}$). 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: $(R_{t-1,t+1})^2 = (R_{t-1,t})^2 + (R_{t,t+1})^2 + 2(R_{t-1,t})(R_{t,t+1})cos(180^0 - \theta_t)$, where $(R_{t,t'})$ is the Euclidean distance between positions $(x,y)_t$ and $(x,y)_{t'}$.

Figure: Turn angle and Step Size calculation.

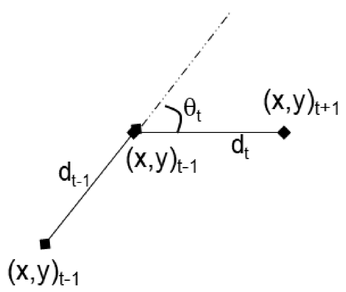
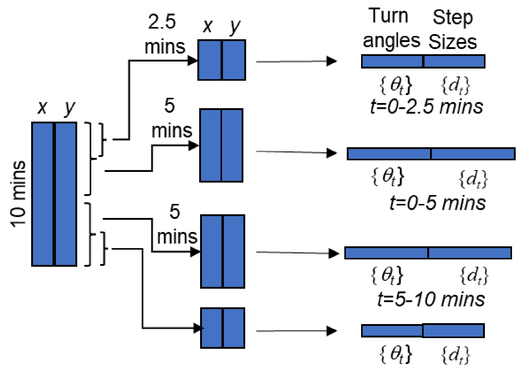


Figure: An illustration demonstrating extraction of turn angle and step size features from a trajectory is shown.

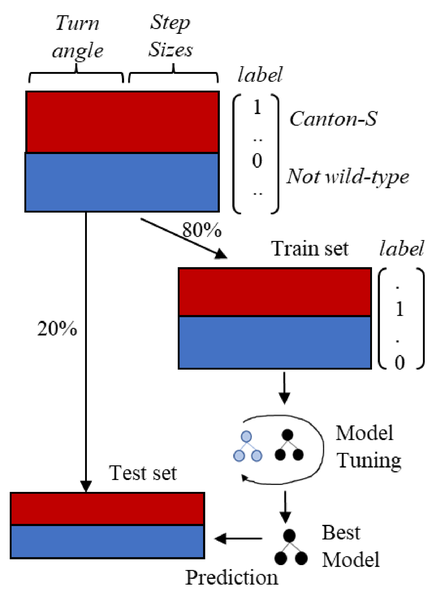


The goal was use supervised learning to test whether the genotype of a fruit fly g_i based on the feature vector z_i. In our context, g_i is the genotype of the fly used in the i^{th} experiment, z_i contains turn angles and step sizes of the i^{th} fly at different time points. Since the duration of one experiment was 10 minute (600 time points) long, We also considered four other cases, where only sections of the 10-minute duration was considered: first 2.5 minutes and 5 minutes, last 2.5 and 5 minutes.

Models and Training

The total 529 experiments were split into training (80%) and testing sets (20%). 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). 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. For the Explainable Boosting classifier, the "interpret" python library was used.

Figure: Procedure for model training and testing

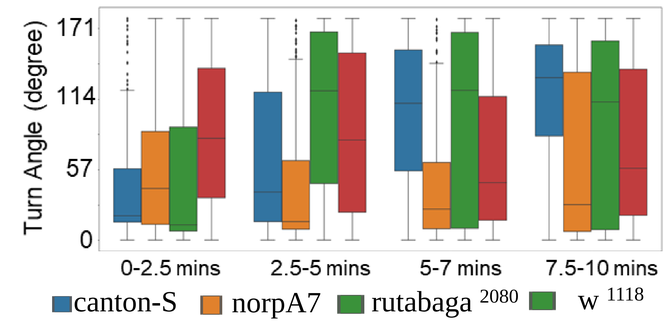


Results & Conclusions

Turn angle and step size are dependent on fly genotype as well as on time

The turn angles for mutant flies *norpA*^{7} and w^{1118} were not significantly different across the four-time sections of the 10-minute duration. For the other two genotypes (*canton-S* and *rutabaga*), the turn angles were significantly different across the time sections. At any time, section during the 10-minute duration, the turn angles and step sizes are significantly affected by the genotype of the fly.

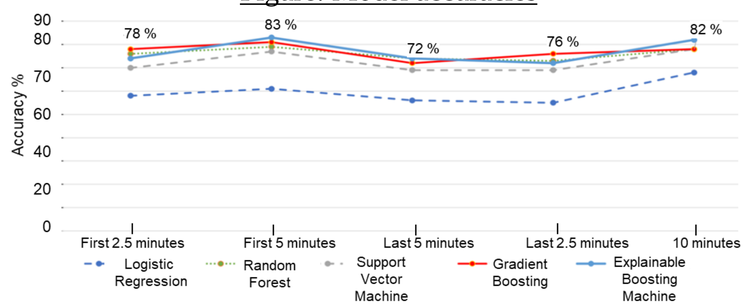
Figure: Box plot of turn angle w.r.t time and genotype



Explainable Boosting Classifier model predicted fruit fly genotype accurately

When the first 5 minutes of the trajectories were considered, Explainable Boosting Classifier achieved the highest accuracy of 83% followed by Gradient Boosting with 80%.

Figure: Model accuracies



Turn angles are better predictors for fruit fly genotype

When the first 5 minutes were considered, turn angles at turn angles during the first one minute have the highest importance scores. The importance scores of turn angles decreased as time increases following a logarithmic

Figure: Important scores of turn angles and step size in predicting genotype