A machine vision based frailty index for mice

Leinani E. Hession[*,1], Gautam S. Sabnis[*,1], Gary A. Churchill[1, **], and Vivek Kumar[1, **]

[*]Equal Contribution
[1]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609
[**]Corresponding Authors: Gary.Churchill@Jax.org, Vivek.Kumar@Jax.org

May 7, 2021

1 Abstract

Chronological aging is uniform, but biological aging is heterogeneous. Clinically, this heterogeneity manifests itself in health status and mortality, and it distinguishes healthy from unhealthy aging. Clinical frailty indexes (FIs) serve as an important tool in gerontology to capture health status. FIs have been adapted for use in mice and are an effective predictor of mortality risk. To accelerate our understanding of biological aging, high-throughput approaches to pre-clinical studies are necessary. Currently, however, mouse frailty indexing is manual and relies on trained scorers, which imposes limits on scalability and reliability. Here, we introduce a machine learning based visual frailty index (vFI) for mice that operates on video data from an open field assay. We generate a large mouse FI datasets comprising 256 males and 195 females. From video data on these same mice, we use neural networks to extract morphometric, gait, and other behavioral features that correlate with manual FI score and age. We use these features to train a regression model that accurately predicts frailty within 1.03 ± 0.08 ($3.9\% \pm 0.3\%$) of the pre-normalized FI score in terms of median absolute error. We show that features of biological aging are encoded in open-field video data and can be used to construct a vFI that can complement or replace current manual FI methods. We use the vFI data to examine sex-specific aspects of aging in mice. This vFI provides increased accuracy, reproducibility, and scalability, that will enable large scale mechanistic and interventional studies of aging in mice.

2 Introduction

Aging is a terminal process that affects all biological systems. Biological aging– in contrast to chronological aging– occurs at different rates for different individuals. In humans, growing old comes with increased health issues and mortality rates, yet some individuals live long and healthy lives while others succumb earlier to diseases and disorders. More precisely, there is an observed heterogeneity in mortality risk and health status among individuals within an age cohort [1, 2]. The concept of frailty is used to quantify this heterogeneity and is defined as the state of increased vulnerability to adverse health outcomes [3]. Identifying frailty is clinically important as frail individuals have increased risk of diseases and disorders, worse health outcomes from the same disease, and even different symptoms of the same disease [2].

The frailty index (FI) is an invaluable and widely used tool which outperforms other methods to quantify frailty [4]. In this method, an individual is scored on a set of age-related health deficits to produce a cumulative score. Each deficit must have the following characteristics: they must be health related, they must increase in the population with age, and they must not saturate in the population too

early [5]. The presence and severity of each health deficit is scored as 0 for not present, 0.5 for partially present, or 1 for present. A compelling finding of FIs is that the exact health deficits scored can vary between indexes but still show similar characteristics and utility [5]. That is, two sufficiently large FIs with different numbers and selections of deficits scored would still show a similar average rate of deficit accumulation with age and the same submaximal limit (the highest FI score observed). More importantly, both FIs would predict an individual's risk of adverse health outcomes, hospitalization, and mortality. This feature of FIs is advantageous as researchers can pull data from varied large health databases, aiding in large-scale studies. It also suggests that frailty is a legitimate phenomenon and that FIs are a valid measure of biological aging. Different people age not only at different rates but in different ways; one person may have severe mobility issues but have a sharp memory, while another may have a healthy heart but a weak immune system. Both may be equally frail, but this is only made clear by sampling a variety of health deficits. Indeed, FI scores outperform other developed measures including molecular markers and frailty phenotyping at predicting mortality risk and health status [4, 6, 7].

FIs have been adapted for use in mice using a variety of both behavioral and physiological measures as index items [2, 4, 8]. Mouse FI shows many of the characteristics of human FIs, including a submaximal limit and a strong correlation with mortality [9]. Mouse FIs have been successfully used to evaluate a variety of aging interventions [10] and for construction of models of chronological age and mortality [4]. Unlike human FIs, the majority of mouse frailty indexing has been performed using the same set of health-deficits as Whitehead et al. [2], although some studies have substituted a small number of mouse FI items for ones that better fit their specific strain or experiment [10].

The successful creation of the mouse FI is a major step forward in aging research, particularly for interventional studies that follow health outcomes over long periods of time and are often carried out by multiple labs or at disparate locations. However, because conducting mouse FI requires trained individuals for manual scoring, it often limits the scalability of the tool; FI scoring of hundreds of mice is a labor-intensive task. While human studies can draw from large health databases and have sample sizes in the thousands, mouse studies that employ FIs are much smaller due to the low-throughput nature of the scoring [9]. Furthermore, since many of the FI metrics require some level of subjective judgment, there are concerns about scorer-based variability and reproducibility [10]. In fact, the question of inter-scorer reliability is specifically mentioned as a future research area for mouse FI [2]. For instance, recent work has shown that the professional background of scorers affects FI scores; specifically, inter-scorer reproducibility was poor between animal technicians and research scientists [11]. Although most studies show moderate to high inter-scorer reliability between scorers, they strongly emphasize the importance of inter-scorer discussion in obtaining consistent results, which is not always feasible in multisite or long-term studies [12]. Thus, although the FI is an extremely useful tool for aging research, an increase in its scalability, reliability, and reproducibility through automation would enhance its utility.

Towards this end, we developed an automated visual FI using video of mice in the open field. The open field is one of the oldest and most widely used assays for rodent behavior [13]. Commonly, measures like total distance travelled, location of activity (thigmotaxis), grooming bouts, urination, and defecation have been used to infer the behavioral state of the animal [14]. However, advances in machine learning techniques have greatly expanded the types of metrics that can be extracted from the open field assay beyond the traditional metrics of hyperactivity and anxiety [15, 16]. These advances are largely due to discoveries in the computer vision and statistical learning field [17–22]. We along with a number of other groups have applied these new methods to animal behavior analysis. Our group has developed methods for image segmentation and tracking in complex environments [23], action detection [24] and pose-based gait and whole body coordination measurements in the open field [25]. These and other highly sensitive methods have advanced animal behavior extraction [15,

26, 27].

Our goal was to develop an efficient scalable method to determine frailty in the mouse using computer-vision based features. We hypothesized that biological aging produces changes in behavior and physiology that are encoded in video data, i.e. we can visually determine the frailty of an animal based on their open field behavior. Additionally, sex differences in frailty are less well understood in mice because most mouse frailty experiments have studied only males [9, 28]. Given that many age related changes are known to be sex-specific in humans [29–31], this is a large blind spot in mouse frailty research. Therefore, we generate one of the largest mouse FI dataset consisting of both males and females. We extract sensitive measures of mouse movement (gait and posture), morphometric, and other behaviors using machine learning methods. We use these features to construct a vFI assay that has high prediction accuracy. Through modeling we also gain insight into which video features are important to predict FI score across age and frailty status. Our automated vFI will increase efficiency and accuracy for large scale studies that explore mechanisms and interventions of aging.

3 Results

3.1 Data collection

Our study design evaluated 451 individual C57BL/6J mice (256 males, 195 females), with 117 mice repeated in a second round of testing 5 months later, resulting in a 568 data points for mice ranging from 8 to 148 weeks of age (Figure 1A). Top-down video of each mouse in a one hour open field session was collected according to previously published protocols [16, 23] (see Methods, Figure 1A, and supplementary Video 1 and 2 as examples of a young and old mouse). Following the open field, each mouse was scored using a standard mouse frailty indexing by a trained expert from the Nathan Shock Center for Aging to assign a manual FI score [32]. Over the course of the data collection, 4 different scorers conducted the manual FI. The open field video was processed by a tracking network and a pose estimation network, to produce a track, an ellipse-fit, and a 12-point pose of the mouse for each frame [23, 25]. These frame-by-frame measurements were used to calculate a variety of per-video features. These features included traditional open field measures such as anxiety, hyperactivity [23], neural network-based grooming [24], and novel gait measures [25]. All features are defined in Section 3.2 and Supplementary Table S1. The per-video features for each mouse were used as features in an array of machine learning models, including penalized linear regression (LR*) [33], random forest (RF) [34], support vector machine (SVM) [35], and extreme gradient boosting (XGB) [36]. The manual FI scores were used as the response variables for the models. As expected, mean FI score increases with increasing age (Figure 1B). Heterogeneity of FI scores (shown by the standard deviation bars) also increases with age. We find a sub-maximal limit of FI score slightly below 0.5 for our data, which falls within a range of submaximal limits shown in mice [2, 9]. These results show that our FI data are typical of other mouse data and mirrors the characteristics of human FIs with the increase in average FI scores and heterogeneity of FI scores with age [9]. Visual inspection of the data indicated that there may be a scorer-dependent effect on the manual FI. For instance, Scorer 1 and 2 tend to generate high and low frailty scores, respectively. We investigated the effect of scorer using a linear mixed model with scorer as the random effect and found 35% of the variability (RLRT $= 66.41$, $p < 2.2e^{-16}$) in manual FI scores could be accounted for by scorer (Figure 1C). Restricted likelihood-ratio test (RLRT) [37] provided strong evidence of scorer (random) effect with non-zero variance. This suggests variability between scorers is an important source of variability in our data and should be adjusted prior to modeling.

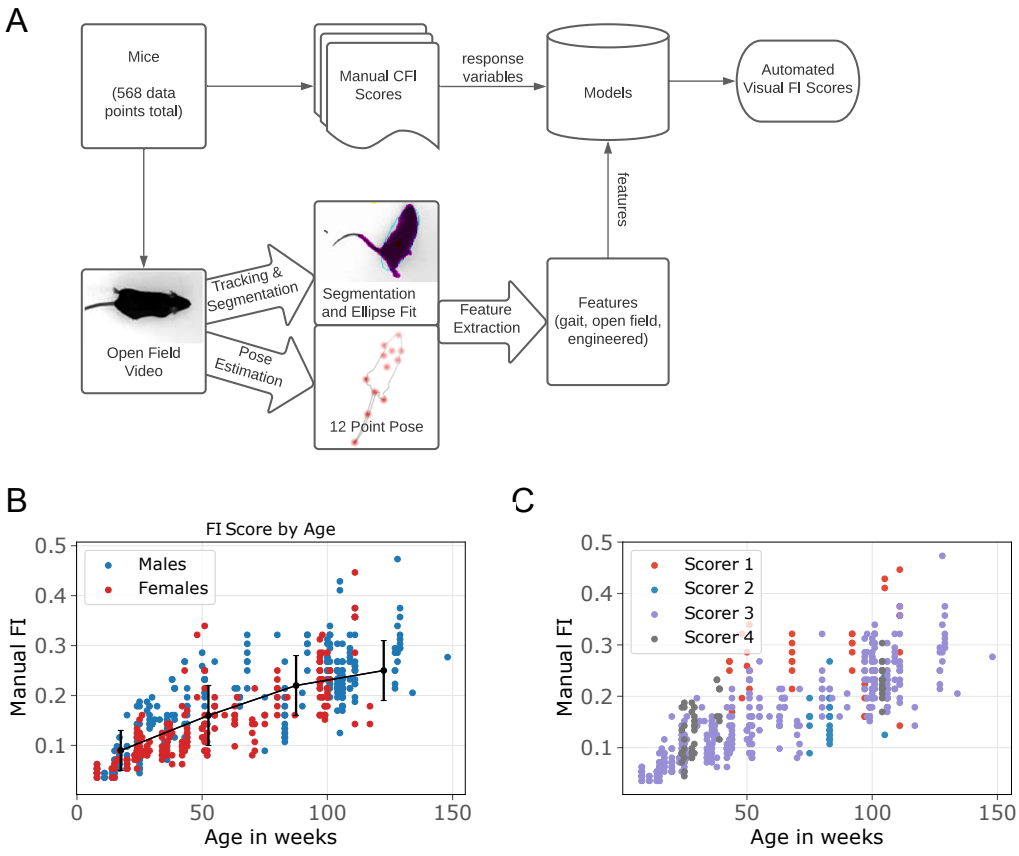

Figure 1: General approach to build a visual frailty index. (A) Pipeline for automated visual frailty index (vFI). Top-down videos of the open field for each mouse are processed by a tracking and segmentation network and a pose estimation network. The resulting frame-by-frame ellipse-fits and 12 point pose coordinates are further processed to produce per-video metrics for the mouse. The mouse is also manually frailty indexed to produce a FI score. The video features for each mouse are used to model its FI score. B) A scatter plot of FI score by age. The black line shows a piece-wise linear fit to the data and the error bars are the standard deviations. C) Scatter plot of FI score by age colored by scorer.

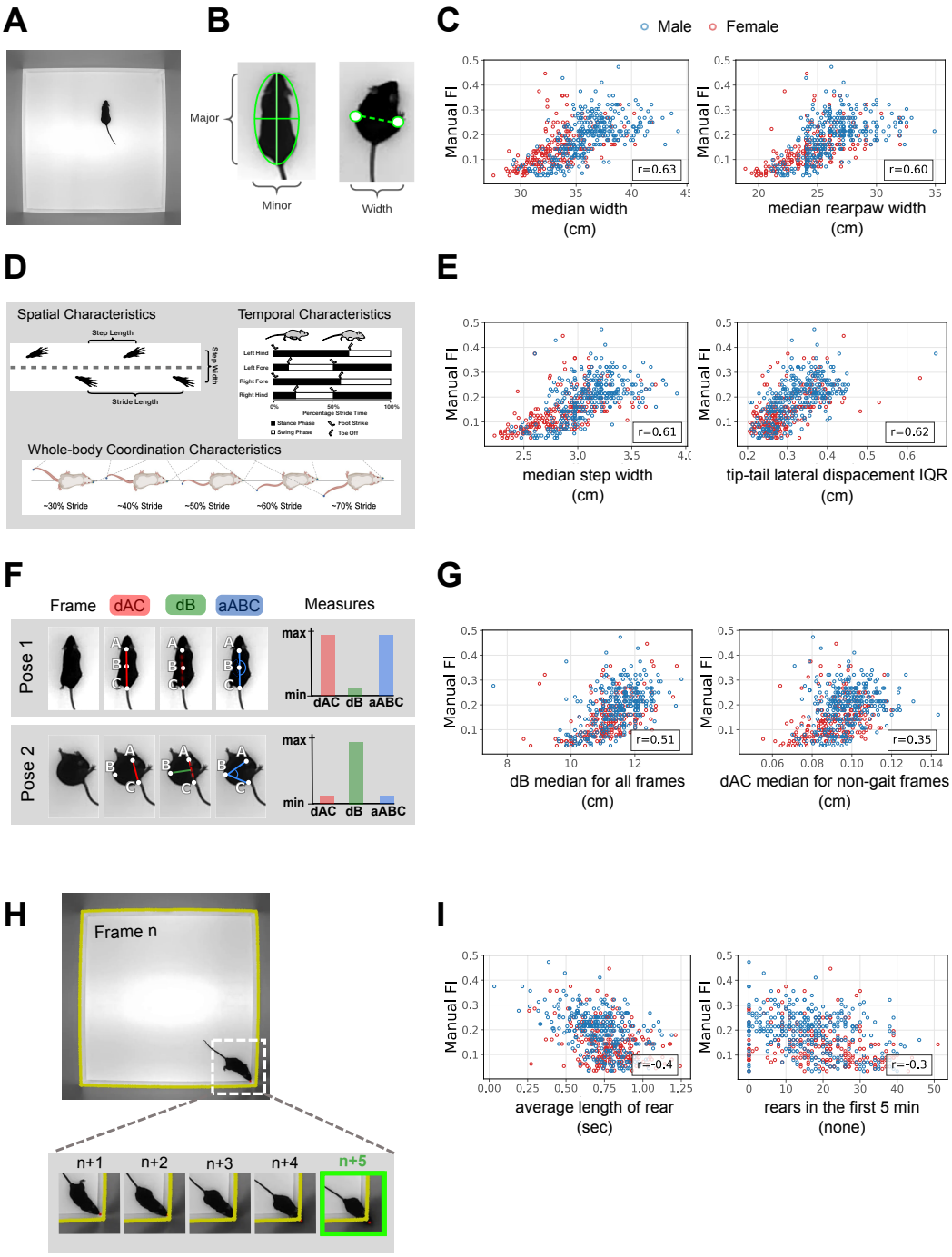

Figure 2: Sample features used in the vFI (A) Single frame of the top-down open-field video. (B) Morphopemetric features from ellipse-fit and rear-paw distance measure performed on the mouse frame by frame. The major and minor axis of the ellipse fit are taken as the length and width respectively. (C) The median ellipse-fit width and the median rear-paw distance taken over all mouse frames are highly correlated with FI score. (D) Spatial, temporal, and whole-body coordination characteristics of gait used to create metrics ([25]). (E) The median step width and the inter-quartile range of tip-tail lateral displacement taken over all strides for a mouse are highly correlated with FI score. (F) Spinal mobility measurements taken at each frame. dAC is the distance between point A and C (base of head and base of tail respectively) normalized for body length, dB is the distance of point B (mid-back) from the midpoint of the line AC, and aABC is the angle formed by the points A, B, and C. When the mouse spine is straight, dAC and aABC are at their maximum value while dB is at its minimum. When the mouse spine is bent, dB is at its maximum value while dAC and aABC are at their minimum. See Supplementary Video 3.(G) The median of dB taken over all mouse frames shows a strong correlation with FI score, and the median dAC taken only over frames where the mouse is not in gait shows a moderate correlation with FI score. (H) Wall rearing event. The contour of the walls of the open field are taken and a buffer of 5 pixels is added (yellow line), marking a threshold. The nose point of the mouse is tracked at each frame. A wall rearing event is defined by the nose point fully crossing the wall threshold. See Supplementary Video 4. (I) The mean length of rearing event taken over all rearing events and the number of rears in the first 5 minutes of the open-field video show moderate correlation with FI score.

3.2 Feature extraction

The frame-by-frame segmentation, ellipse fit, and 12-point pose coordinates were used to extract per-video features [23–25]. All extracted features with explanation and source of the measurements can be found in Supplementary Table S1. Overall, there is a very high correlation between median and mean video metrics (Fig S1A,B). We decided to use only medians in modeling for two reasons: medians tend to have higher correlation with FI score than means, and medians are more robust to outlier effects than means. Likewise, we decided to use inter-quartile ranges when available and not standard deviations as features in the models, as inter-quartile range tends to be more robust to outliers than standard deviations.

We first looked at metrics taken in standard open field assays such as total locomotor activity, time spent in the periphery vs. center, and grooming bouts (Figure 2A). All standard open-field measures showed low correlation with both FI score and age (Supplementary Table S2 and S3).

In addition to the existing features, we designed a set of features that we hypothesized may correlate with FI. These include morphometric features that capture the animals shape and size, as well as behavioral features that are associated with flexibility and vertical movement. Changes in body composition and fat distribution with age are observed in humans and rodents [38]. We hypothesized that body composition measurements may show some signal of aging and frailty. We took the major and minor axes of the ellipse fitted to the mouse at each frame as an estimated length and width of the mouse respectively (Figure 2B). The distance between the rear paw coordinates for each frame were taken as another width measurement closer to the hips. The means and medians of the ellipse width, ellipse length, and rear paw width over all frames were used as per-video metrics. Many of these morphometric features showed high correlations with FI score and age (Supplementary Table S2 and S3), for example, specifically median width and median rear paw width had correlations of r = 0.63 and 0.60, respectively (Figure 2C).

Changes in gait are a hallmark of aging in humans[39, 40] and mice [41, 42]. Recently we established methods to extract gait measures from freely moving mice in the open field [25]. We carried out similar analysis to explore age-related gait changes in the current cohort of mice (Figure 2D, E). Each stride is analyzed for its spatial, temporal, and whole-body coordination measures (Figure 2D), resulting in an array of measures of which the medians over all strides for each mouse are taken. We also looked into intra-mouse heterogeneity of gait features using standard deviations and inter-quartile range over all strides for each mouse. Many of these calculated metrics show a high correlation with FI score and age (Supplementary Table S2 and S3), for example, median step width and tip-tail lateral displacement interquartile range (r=0.61 and r=0.61 , respectively) (Figure 2E).

We next investigated the bend of the spine throughout the video (see Supplementary Video 3). We hypothesized that aged mice may bend their spine to a lesser degree, or less often due to reduced flexibility or spine mobility. This change in flexibility can be captured by the pose estimation coordinates of three points on the mouse at each video frame: the back of the head (A), the middle of the back (B), and the base of the tail (C). At each frame, the distance between points A and C normalized for mouse length (dAC), the orthogonal distance of the middle of the back B from the line (dB), and the angle of the three points (aABC) are calculated (Figure 2F). For each of the three per-frame measures (dAC, dB, and aABC) a mean, median, standard deviation, minimum, and maximum are calculated per video for all frames and for non-gait frames (frames where the mouse is not in stride). We find some moderately high correlations showing relationships between spinal bend and FI score which contradict our hypothesis (Supplementary Table S2 and S3); while we would expect dB median and dAB median (for non-gait frames) to decrease with age, we find that they increase (r=0.51 and 0.35, respectively) (Figure 2G). One possible reason for this result is that very frail mice were spending more time grooming. However, neither grooming bouts or grooming sec show a relationship with FI

score or dB median. Another possibility is that high frailty mice walked less and spent more time curled up. This does not seem to be the case either, as there is almost no relationship between either stride count or distance travelled and FI score or dB median. High frailty mice may also have higher dB medians due to body composition, as dB median has a correlation of 0.496 with body weight. It is important to note that these bend metrics cast a wide net; they are an inexpensive and general account of all the activity of the spine during the one hour open field. Thus, these measures may be capturing the interaction between body composition and behavior.

The final group of metrics looked at occurrences of rearing supported by the wall (Figure 4H, Supplementary Video 4). We hypothesized that frailer mice may rear less. The edges of the open field were taken and a buffer of 5 pixels added as a boundary. We took the frames in which the nose coordinates of the mouse cross that boundary as instances of rearing. From these heuristic rules we are able to determine the number of rears and the average length of each rearing bout (Table S1). We find that there is moderate signal for frailty in some metrics related to rearing bout and length, specifically average length of rear and rears in the first 5 minutes (r = 0.4 and 0.3, respectively, Figure 2I).

Interestingly, the correlations with age were generally slightly higher than FI score (Supplementary Table S2 and S3). This may be due to how mice become frail in different ways; one mouse can have high frailty but no dysfunction in their stride width, while on average older mice, regardless of their frailty, may have more stride width changes.

3.3 Sex differences in frailty

To visualize sex differences in frailty, we stratified the FI score data into four age groups and compared the boxplots for each age group between males and females (Figure 3A). The oldest age group included only 9 females compared to 81 males. The range of females frailty scores for each age group tends to fall slightly lower than males except for the oldest age group. The middle two age groups show highly significant differences in distribution between males and females.

Comparisons between the correlations of male and female FI item scores with age (Figure 3B), show an overall high correlation (r=0.91). The average difference between male and female correlations of FI index items with age was 0.08, but there were a few index items showing notable differences. Alopecia and Tremor have a higher correlation with age for females (by 0.33 and 0.22 respectively), while Loss of fur color and Coat condition are higher for males (by 0.28 and 0.27 respectively) (Supplementary Table S4).

The correlations of male and female video features with both FI score and age were also high (r=0.91 and r=0.92 respectively), with an average difference between male and female correlations of video metrics with both FI score and age of 0.10 (Supplementary Table S2 and S3). In both FI score and age, the video features with the highest sex differences were gait measures related to stride and step length, base tail lateral displacement, and tip tail lateral displacement. The highest differences were the correlations between median base tail lateral displacement and age (difference of 0.48) and median tip tail lateral displacement and age (0.42), for which females tended to have a higher correlation to both FI score and age. For the metrics related to stride length and step length (median step length, median stride length, and step length standard deviation and interquartile range), males had a higher correlation to FI score and age than females.

These gait features with the highest differences in correlation between sexes seemed to show a pattern; lateral displacement measures showed a higher signal for frailty and age in females while stride length and step length measures showed a higher signal for males.

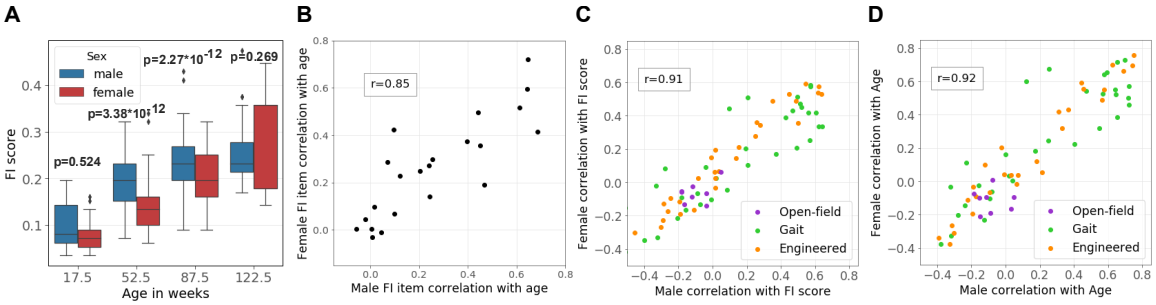

Figure 3: Comparison of male and female FI metrics. (A) The distribution of FI scores for males and females when the data are split into 4 age groups of equal range. The x-ticks represent the midpoint of each age group range. Significant difference in the distribution of male and female scores for that age group determined by the Mann-Whitney U test. (B) Pearson correlations of FI items with age for males compared to females. (C) Pearson correlations of video metrics with FI score for males compared to females. (D) Pearson correlations of video metrics with age for males compared to females.

3.4 Modeling video-generated features to predict age and frailty index

In order to determine how accurately we can predict an animal's age and FI score (response variables) from the video data, we used median and inter-quartile range metrics calculated from the video frames as input features (covariates). We adjusted for the tester effect in FI scores as previously described using a linear mixed model (LMM) before predictive modeling. We compared accuracy in predicting response variables (Figure 4 A,C) from the features (Figure 4 B,D) using four models - penalized linear regression (LR*) [33], support vector machine (SVM) [35], random forest (RF) [34], and extreme gradient boosting (XGB) [36]. The data consisted of 451 unique mice, with 116 mice repeated in a second round of testing, giving rise to 568 observations on 38 covariates belonging to three distinct types (Figure 4 B,D). We split the data randomly into two parts, train (80%) and test (20%), and ensured that the repeat measurements from the same animal belonged to either the training or the test data and not both. We used the training data to estimate and tune the models' hyper-parameters using 10-fold cross-validation; the test set served as an independent evaluation sample for the models' predictive performance. We performed 50 different splits on the data to allow for a proper assessment of uncertainty in our test set results. We selected the random forest regression model for predicting age on unseen future data due to its superior performance over other models with a lowest median absolute error (MAE) ($p < 2.2e^{-16}$, $F_{3,147} = 108.24$), root-mean-squared error (RMSE) ($p < 2.2e^{-16}$, $F_{3,147} = 64.81$), and highest R^2 ($p < 2.2e^{-16}$, $F_{3,147} = 50.69$) when compared using repeated-measures ANOVA (Figure 4E). Similarly, the random forest regression model for predicting frailty index on unseen future data performed better than all other models, with a lowest median absolute error (MAE) ($p < 4.3e^{-12}$,$F_{3,147} = 22.60$), root-mean-squared error (RMSE) ($p < 1.86e^{-12}$, $F_{3,147} = 23.44$), and highest R^2 ($p < 3.6e^{-12}$, $F_{3,147} = 22.79$) (Figure 4F). Thus, given new video-generated features as input to the random forest model, we can predict the animal's age to be within 13.19 ± 1.08 weeks of the actual age. We can also predict the FI score to be within 1.03 ± 0.08 ($3.9\% \pm 0.3\%$) of the actual frailty index, thereby demonstrating the robustness of the model. We conclude that frailty and age information is encoded in video data features that we have designed and can be successfully used to construct a vFI.

3.5 Quantifying uncertainty in frailty index predictions

In addition to quantifying an average accuracy, we investigated the prediction errors more closely within our data set. We quantified the prediction error by providing prediction intervals (PIs) that give a range of values, containing the unknown age and FI score with a specified level of confidence,

based on the same data that gives random forest point predictions [43]. One existing approach for obtaining random forest-based prediction intervals involves modeling the conditional distribution of FI given the features using generalized random forests [44, 45]. For animals in the test set, we use generalized random forests based on quantiles to provide the point predictions of the FI score (Age resp.) and prediction intervals, which give a range of FI (Age resp.) values that will contain the unknown FI scores (resp. Age) with 95% confidence (Figure 4J,I). The average PI width for all test animals' predicted FI score is 5.72 ± 1.49 (resp. 80.29 ± 16.8 for predicted Age), while the PI lengths range from 2.3 to 8.5 (resp. 28 to 114 for Age), highlighting that the widths of the PIs are animal and age-group specific. We plotted a smoothed regression fit for PIs' width versus age that indicated the widths increased with the animal's age (Figure 4G,H). The variability of 95% PI widths (Figure 4G,H right panels) showed higher variability for animals belonging to the middle age groups (labeled M in pink). We went beyond simple point predictions by providing prediction intervals (PIs) of the frailty index to quantify our predictions' uncertainty. It allowed us to pinpoint the FI score and age with higher accuracy for some animals than others.

3.6 Feature Importance for Frail and Healthy Animals

A useful visual FI (vFI) should depend on several features that can capture the animal's inherent frailty and simultaneously be interpretable. We took two approaches to identify features important for making vFI predictions using the trained random forest model: (1) feature importance and (2) feature interaction strengths. The feature importance provides a measure of how often the random forest model uses the feature at different depths in the forest. A higher importance value indicates that the feature occurs at the top of the forest and is thus crucial for building the predictive model. For the second approach, we derived a total interaction measure that tells us to what extent a feature interacts with all other model features.

For the feature importance approach, we obtained a more complete picture of the feature importance by modeling three different quantiles of the FI score's conditional distribution. The three quantiles represent three frailty groups: low frail (Q1), intermediate frail (M), and high frail (Q3) animals. We hypothesized that different sets of features are crucial for animals belonging to different frailty groups. Indeed, dB and step length features were crucial in animals belonging to Q1 (green) and Q3 (red) quantiles (Figure 5A). In contrast, features such as length, rear paw, and step width were important for lower frailty animals. Similarly, step width, tip tail LD, and width were critical for animals with an FI score close to M (blue).

For the feature interaction strength approach, we used the H-statistic [46] as an interaction metric that measures the fraction of variability in predictions explained by feature interactions after considering the individual features (Figure 5C). For example, we can explain 14% of the prediction function variability due to interaction between tip tail LD and other features after considering the individual contributions due to tip tail LD and other features. We can explain about 13% (resp. 12%) of the prediction function variability due to interaction between width (resp. step length) and other features. We dove deeper and inspected all the two-way interactions between tip tail LD and the other features (results not shown). We found strong tip tail LD interactions with width, stride length, rear paw, and dB of the animal.

Both feature importance and feature interaction strengths informed us that the trained random forest for vFI depends on several features and their interactions. However, they did not tell us how the vFI depends on these features and how the interactions look. We used the accumulated local effect (ALE) plots [47] that describe how features influence the random forest model's vFI predictions on average (Figure 5B). For example, an increasing tip tail lateral displacement positively impacts (increases) the predicted FI score for animals in intermediate and high frail groups (blue and green). Similarly, an increasing rear paw measure positively affects the predictions - the impact is most visible for animals

in the high frail group. Animals with larger widths positively affect predictions; larger step widths and dBs positively impact model predictions. Thus the ALE plots for important features provide clear interpretations that are in agreement with our intuition. We explored the ALE second-order interaction effect plot for the step length1-step width (Figure 5D) and Length-Width (Figure 5E) predictors. It revealed the two features' additional interaction effects and did not include the main features' marginal effects. Figure 5D revealed an interaction between step width and step length: larger step widths and step lengths increased the predicted FI scores. Similarly, larger widths (36 - 44 cms) and lengths (52 - 60 cms) positively impacted the average FI scores predictions.

To summarize, we established vFI's utility by demonstrating its dependence on several features through marginal feature importance and feature interactions. Next, we used the ALE plots to understand the effects of features on the model predictions, which helped us relate the black-box models' predictions to some of our video-generated features. Opening the black-box model was an essential final step in our modeling framework.

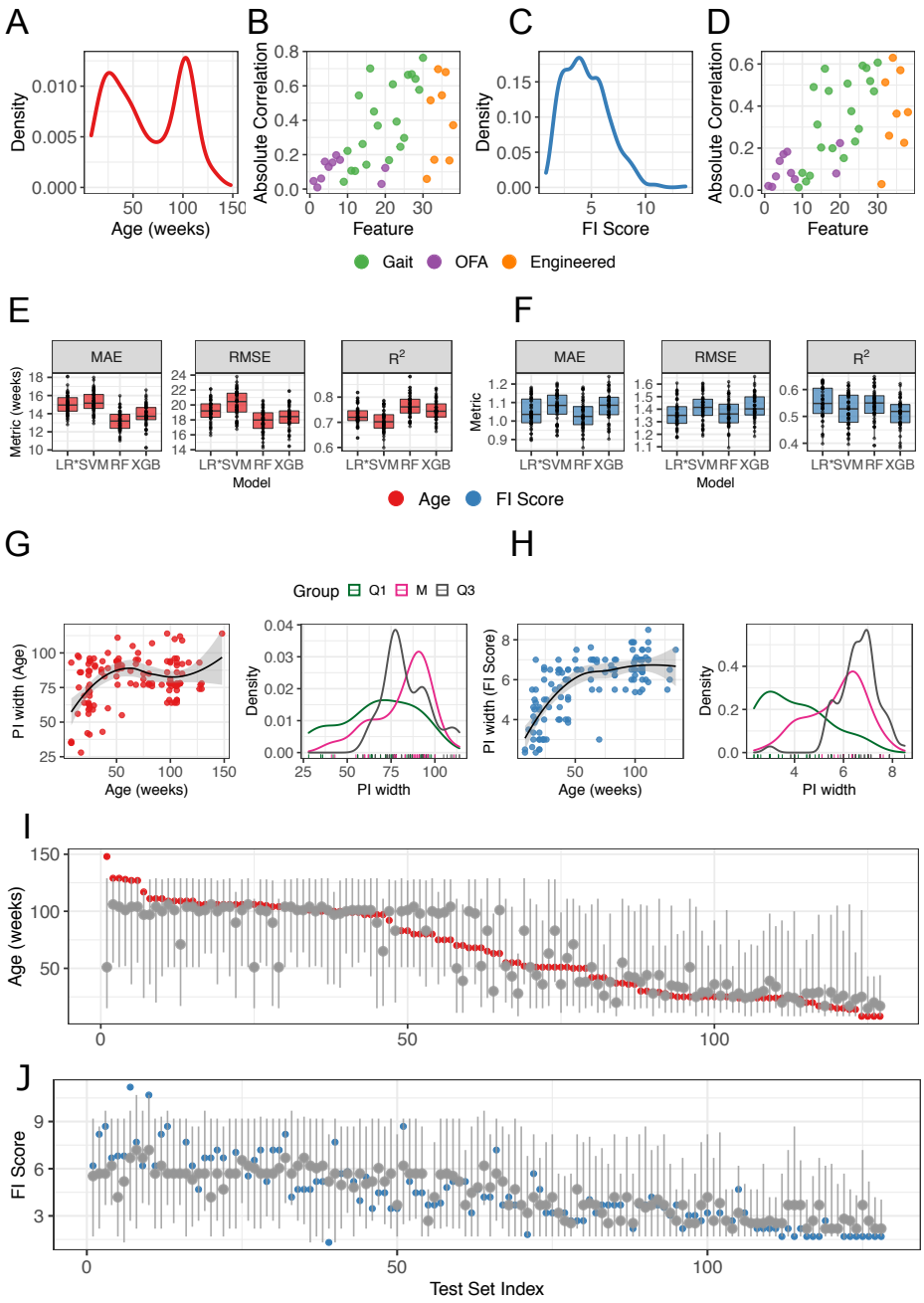

Figure 4: Predicting Age and Frailty from video features. (A) The distribution of Age (weeks) across 451 animals. (B) Absolute correlations of different types of features (Gait, OFA, Engineered) with Age. (C) The distribution of FI scores (range 0-27) across 451 animals. (D) Absolute correlations of different types of features (Gait, OFA, Engineered) with FI scores. (E, F) Comparison among four models (LR*, SVM, RF, XGB) for predicting Age (E) and FI score (F) from video features in terms of median absolute error (MAE), root-mean-squared error (RSME), and R^2 shows that RF outperforms other models. (G, H - 1st panel) We calculate the uncertainty in predicting Age (G) and FI score (H) and plot it as a Function of Age. The black curve shows the loess fit. These plots show less uncertainty in predicting Age and FI scores for very young/old animals. (G, H - 2nd panel) We plot the distributions of prediction interval (PI) widths and find that the PI widths for predicting Age are wider (increased uncertainty in predictions) for animals belonging to the middle age group (M). Similarly, the PI widths for predicting FI scores increase with Age in our data. (I, J) 95% prediction intervals (PIs) (gray lines) quantifying uncertainty in point estimates/predictions (gray dots). There is one interval per test animal, and approximately 95% of the PI intervals contain the correct Age (red dots) and FI scores (blue dots).

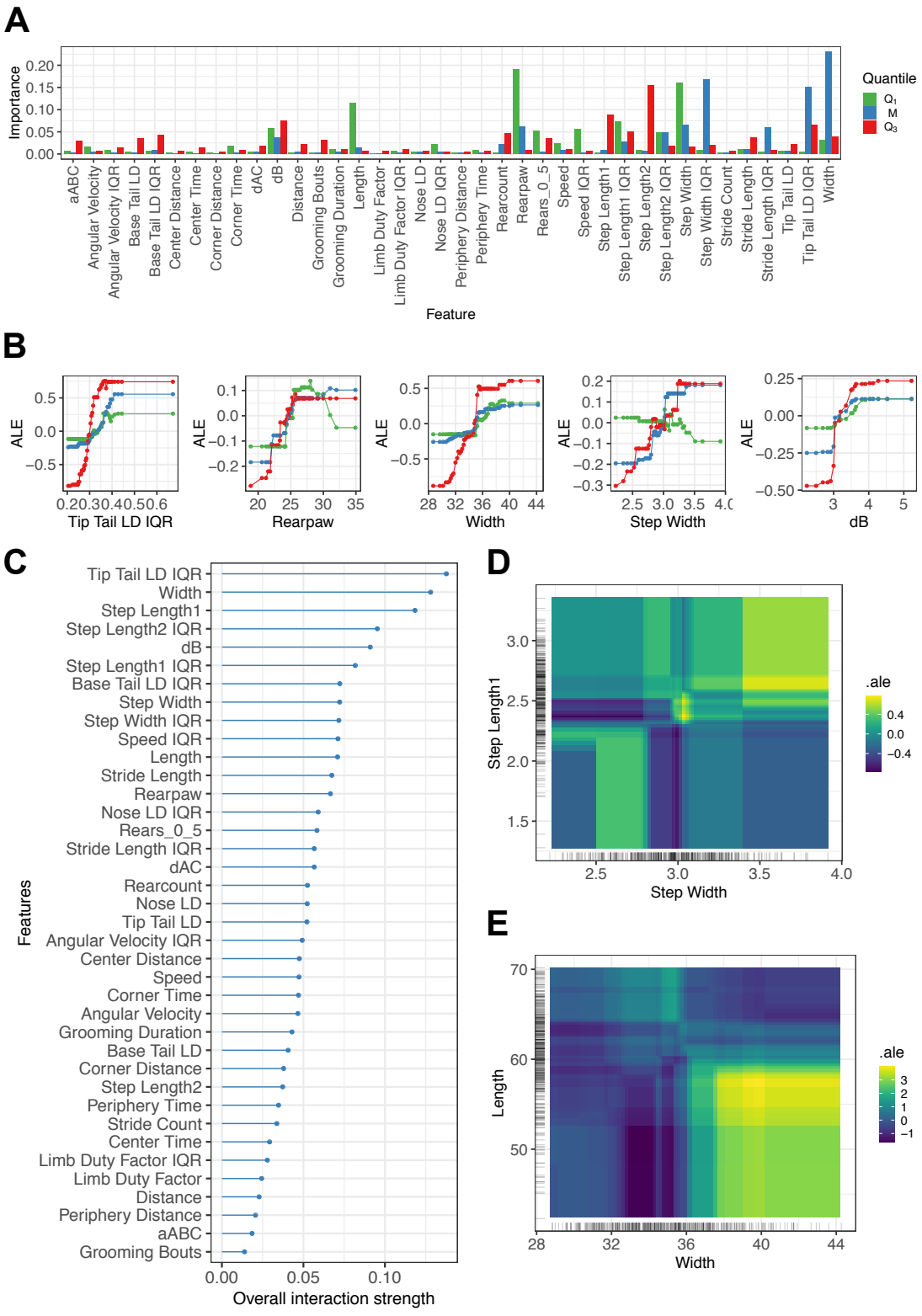

Figure 5: Quantile regression modeling of vFI using Generalized Random Forests. (A) Variable importance measures for three quantile random forest models (lower tail - $Q_{.025}$, median - $Q_{.50}$,upper tail - $Q_{.975}$). Animals in lower and upper tail correspond to animals with low and high frailty scores respectively. (B) Marginal ALE plots show how important features influence the predictions of our models on average. For example, the average predicted FI score rises with increasing step width, but falls for values greater than 3 in animals belonging to lower and upper tail. (C) A plot showing how strongly features interact with each other. (D,E) ALE second-order interaction plots for step width and step length1 (E: Width and Length) on the predicted FI score. Lighter shade indicates an above average and darker shade a below average prediction when the marginal effects from features are already taken into account. Plot D (resp. E) reveals a weak (resp. strong) interaction between step width and step length1 (resp. width and length). Large step width and step length1 increases the predicted FI score.

4 Discussion

The mouse FI is an invaluable tool in the study of biological aging. Here, we seek to extend it by producing an automated visual frailty index (vFI) using video-generated features to model FI score. This vFI offers a reliable high-throughput method for studying aging. We generate one of the largest frailty data sets for the mouse with associated open field video data. We use computer vision techniques to extract behavioral and morphometric features, many of which show strong correlations with aging and frailty. We also analyze sex-specific aging in mice. We then train machine learning classifiers that can accurately predict frailty from video features. Through modeling we also gain insight into feature importance across age and frailty status.

The mice were assigned FI scores with a standard manual FI test. Average FI scores and heterogeneity increased with age as expected. Our sub-maximal limit was slightly below 0.5, which falls well within the range of previous research [2, 9]. We find that 35% of the variability in our data set could be accounted for by scorer, indicating the presence of a tester effect. Although previous studies looking at tester effect found good to high inter-reliability between testers in most cases [10], it is known that characteristics of the tester and how the tester handles the mouse can affect FI scores [11]. Therefore, reducing the handling of mice may increase reliability of scores.

Top-down videos of mice in the open field were processed by previously trained neural-networks to produce an ellipse-fit and segmentation of the mouse as well as a pose estimation of 12 salient points on the mouse for each frame. These frame-by-frame measures were used to engineer features to use in our models. The first category of features were standard open field metrics such as time spent in the periphery vs center, total distance travelled, and count of grooming bouts. All standard open field metrics had very poor correlation with both FI score and age. These results suggest that standard open field assays meant to measure emotionality are inadequate to study aging.

The next engineered features were morphometric features. In humans, anthropometric measures such as BMI and waist-to-hip ratio are predictors of health conditions and mortality risk [48, 49]. Aging is associated with changes in body measurements and composition, increased visceral fat, and progressive dysfunction and changes in composition of adipose tissue [50]. The effect of aging on body composition in rodent models is less established, and though there are observed changes in body composition and fat mass in growth patterns similar to humans [49, 51]. To investigate the viability of computer vision techniques for morphometric analysis, we take the width of the ellipse fit and the distance between the rear paws coordinates as proxy measures for waist and hip measurements. We find high correlation between these body measurements and both FI score and age, in particular median ellipse width and median rear paw width, suggesting that there is a signal for aging in mouse body composition.

Prevalence of gait disorders increase with age, and specific changes in gait are associated with age-related neuro-degenerative disorders such as Parkinson's and Alzheimer's disease, as well as age-related inflammatory diseases like arthritis [39]. Evaluation of fall-risk and mobility decline using gait metrics in older patients is of clinical concern [39]. Geriatric patients are also shown to have more irregularity in their gait; for example, older adults have been shown to have increased variability in their step width [40]. Using neural networks trained to extract individual strides, we can look at the spatial, temporal, and whole-body coordination characteristics of gait for each mouse. We find many gait metrics with strong correlation to both frailty and age; in particular we find a strong increase in tip-tail lateral displacement heterogeneity, step width, step width variability, and step length variability. Analogous to human data, we find a decrease in stride speed with age, as well as an increase in step width variability [40]. Previous investigations into gait in aging mice have focused on the decline of stride speed and stride time variability [41, 42], but we find varied gait metrics showing high significant correlations to both age and frailty. As gait is thought to have both cognitive and muscular-skeletal

dynamic components, it is a compelling area for aging and frailty research.

Spinal mobility in humans is a predictor of quality of life in aged populations [52] and the mouse is used as a model for the aging human spine [53]. Surprisingly, though some spinal bend metrics show moderately high correlations with FI score, the relationship is the opposite of what we initially hypothesized. As these metrics are a general account of all the activity of the spine during the experiment, they are likely capturing a combination of behaviors and body composition which gives this result. Nevertheless, some of these metrics showed a moderately high correlation with FI score and age and were deemed important features in the model.

Many age-related biochemical and physiological changes are known to be sex-specific [29–31]. Understanding sex differences in the presentation and progression of frailty in mice is crucial for translating pre-clinical results for clinical use. It is of interest to understand how sex characteristics such as hormones and body fat distribution relate to biological aging. In humans, there is a known 'mortality-morbidity paradox' or 'sex-frailty paradox', where women tend to be more frail but paradoxically live longer [28]. In C57BL/6J mice, however, it seems males tend to live slightly longer than females, though there is variability, and females do not seem to paradoxically live longer when frail [28]. We find more males surviving to old age than females, and further, we find females tended to have slightly lower frailty distributions than males of the same age group. These results suggest that in mice, the sex-frailty paradox shown in humans may not exist or may be reversed. We compared the correlations of FI index items to age between males and females and found some sex differences in the strength of correlation for a few of the index items, mostly related to visual fur changes. When comparing the correlations of the video features to age and FI score between males and females, we also find a number of starkly different correlations. Median base tail lateral displacement and median tip tail lateral displacement were both much more strongly correlated with age for females than for males; as female mice age, their tail lateral displacement within a stride tends to increase, while males see almost no change. Males on the other hand show a strong decrease in stride length and a strong increase in step length with age, while females show very little change. Most video features with higher differences are gait-related, with a couple related to spinal bend. These differences in gait with age are a new insight and the exact mechanisms driving them are still unknown. It is important to understand how sex differences in human frailty compare to mouse frailty in order to critically evaluate how results from mouse studies can translate to humans.

Using the video-generated features as input to the random forest model, we could predict the visual frailty index within 1.03 ± 0.08 ($3.9\% \pm 0.3\%$) of the actual frailty index on average of the actual frailty index score, an error within 1 item mis-scored at 1 or 2 items mis-scored at 0.5. Furthermore, we went beyond simple point predictions by providing 95% prediction intervals of the frailty score which contain the unknown continuous univariate FI with a specified level of confidence. We then applied quantile random forests to low and high quantiles of the frailty score's conditional distribution that allowed us to understand the influence of gait, open-field, and engineered features and found that different features affect frail and healthy animals differently.

A limitation of our work is that we did not have large numbers of mice representing the very old age group and that similarly we do not have very frail mice. This reflects ethical decisions about terminating mice that are too sick. In addition, the sex distribution of our data was also quite unbalanced for the older mice. Given that modeling has revealed key differences in the presentation of frailty and age between the very young, middle, and very old mice, it would benefit the model to have a larger sex-balanced and pool of older mice. Since frailty is of specific importance for geriatric patients, a more complete exploration of the oldest mice would have the most clinical significance. Nonetheless, it is of general interest to track the development of frailty across ages, starting from very young mice. Our method relies on locomotion in the open field, which is known to vary across strains [54]. In this trial we use a genetically homogeneous population (C57BL/6J). Given the evidence of a strong

genetic component of aging [55], applying this method to other strains, genetically heterogeneous populations such as Diversity Outcross and Collaborative Cross, may reveal how genetic variation influences frailty and enable the mapping of its genetic architecture. A power of video data is the ability to add features over time. The same videos can be reanalyzed for extraction of new features (behavioral and physiological) enabled by new technology to improve the vFI [56, 57]. Further, as predicting mortality risk is a vital function of frailty, exploring the relationship of our frailty metrics and mortality is important and a starting point for studying lifespan. Additionally, this approach of using computer vision techniques to model frailty could potentially be used in a home cage. Not only would the use of the home cage for the assay further reduce handling and environmental factors in FIs, continuous monitoring of frailty in the home cage environment would ensure the wellness of the mouse and support good husbandry practices.

Overall, our approach has produced some novel insights into mouse frailty and provided a tool which can be used for high-throughput studies. Just as the manual FIs for both humans and mice are based on the principle of sampling many different health deficits spanning different organ systems, by looking at complex behaviors which encompass many aspects of both physiology and cognition along with morphometric features, our vFI achieves the goals of traditional FI methods.

5 Methods

5.1 Mice

C57BL/6J mice were obtained from the Nathan Shock Center at the Jackson Laboratory. The dataset contains 568 trials from 451 individual mice (256 males, 195 females), with 117 mice that were tested twice, between the ages of 8 to 148 weeks of age. The mice were group housed. They were tested in 2 rounds approximately 5 months apart.

5.2 Open Field Assay and Frailty Indexing

The open field behavioral assays were conducted as described in [16, 23]. Mice were shipped from Nathan Shock Center aging colony which resides in different room in the same animal facility at JAX. The aged mice acclimated for 1 week to the Kumar Lab animal holding room, adjacent to the behavioral testing room. During the day of the open field test, mice were allowed to acclimate to the behavior testing room for 30–45 minutes before the start of test. One hour open field testing was performed as previously described [16, 23]. After open field testing mice were returned to the Nathan Shock Center for manual FI. Manual FI was performed by trained experts in the Shock Center within 1 week of the open field assay on each mouse according to previously described protocols[2, 32]. FI testing sheet with all items can be found in Supplementary Materials.

5.3 Video, Segmentation, and Tracking

Our open field arena, video apparatus, and tracking and segmentation networks are as detailed previously [23]. Briefly, the open field arena measures 20.5" by 20.5" with Sentech camera mounted 40 inches above. The camera collects data at 30 fps with a 640x480px resolution. We use a neural network trained to produce a segmentation mask of the mouse to produce an ellipse fit of the mouse at each frame as well as a mouse track.

5.4 Pose Estimation and Gait

The 12-point 2D pose estimation produced using a deep convolutional neural network trained as detailed in (Gait paper). The points captured are nose, left ear, right ear, base of neck, left forepaw, right

forepaw, mid spine, left rear paw, right rear paw, base of tail, mid tail and tip of tail. Each point at each frame has an x coordinate, a y coordinate, and a confidence score. We use a minimum confidence score of 0.3 to determine which points are included in the analysis.

The gait metrics were produced as detailed in Shephard et al. (2020) [25]. Briefly, the stride cycles were defined by starting and ending with the left hind paw strike, tracked by the pose estimation. These strides were then analyzed for several temporal, spatial, and whole-body coordination characteristics, producing the gait metrics over the entire video.

5.5 Open field measures and Feature Engineering

Open field measures were derived from ellipse tracking of mice as described before [23, 24]. The tracking was used to produce locomotor activity and anxiety features. Grooming was classified using a action detection network as previously described [24]. The other engineered features (spinal mobility, body measurements, and rearing) were all derived using the pose estimation data. The spinal mobility metrics used 3 points from the pose: the base of the head (A), the middle of the back (B) and the base of the tail (C). For each frame, the distance between A and C (dAC), the distance between point B and the midpoint of line AC (dB), and the angle formed by the points A,B, and C (aABC) were measured. The means, medians, maximum values, minimum values, and standard deviations of dAC, dB, and aABC were taken over all frames and over frames that were not gait frames (where the animal was not walking). For morphometric measures, we measured the distance between the two rear paw points at each frame and too the means, medians, and standard deviations of that distance over all frames. For rearing, we took the coordinates of the boundary between the floor and wall of the arean (using OpenCV contour) and added a buffer of 4 pixels. Whenever the mouse's nose point crossed the buffer, this frame was counted as a rearing frame. Each uninterrupted series of frames where the mouse was rearing (nose crossing the buffer) was counted as a rearing bout. The total number of bouts, the average length of the bouts, the number of bouts in the first 5 minute, and the number of bouts within minutes 5 to 10 were calculated.

5.6 Modeling

We removed the tester effect from the FI scores using a linear mixed model (LMM) with the lme4 R package [58]. The following model was fit:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \;\; \mu_i \sim P \equiv N(0, \tau^2)$$

where y_{ij} is the jth animal scored by tester i, μ_i is a tester-specific mean, ε_{ij} is the animal-specific residual, σ^2 is the within-tester variance and P is the distribution of tester-specific means. We had 4 testers with different number of animals tested by each tester ie $i = 1, \ldots, 4$. The tester effects, estimated with the best linear unbiased predictors (BLUPs) using restricted maximum likelihood estimates [59] were subtracted from the FI scores of the animals, $\tilde{y}_{ij} = y_{ij} - \hat{\mu}_i$.

We modeled tester-adjusted FI scores, $\tilde{y}_{ij}$, with video-generated features as covariates/inputs using linear regression model with elastic-net penalty [33], support vector machine [35], random forest [34], and gradient boosting machine [36]. We split the data randomly into two parts: train (80%) and test (20%). We used the training data to estimate and tune the models' hyper-parameters using 10-fold cross-validation; the test set served as an independent evaluation sample for the models' predictive performance. We performed 50 different splits on the data to allow for a proper assessment of uncertainty in our test set results. The models were compared in terms of median absolute error (MAE), root-mean-squared-error (RMSE), and R^2. These metrics were compared across the four models using repeated-measures ANOVA through F test with Satterthwaite approximation [60] applied to the test statistic's denominator degrees of freedom.

We obtained the $100(1-\alpha)\%$ out-of-bag prediction intervals $\mathbf{I}_\alpha(\mathbf{X}, \mathbf{C_n})$, where $\mathbf{X}$ is the the vector of covariates and C_n is the training set, via quantile random forests [44] with the grf package [45]. Prediction intervals produced with quantile regression forests often perform well in terms of conditional coverage at or above nominal levels i.e. $\mathbb{P}[\tilde{y} \in \mathbf{I}_\alpha(\mathbf{X}, C_n)|\mathbf{X} = \mathbf{x}] \geq 1-\alpha$ where we set $\alpha = 0.05$.

5.7 Data and Code Availability

All behavioral data will be available in the Mouse Phenome Database, and code and models will be available in Kumar Lab Github account (https://github.com/KumarLabJax and https://www.kumarlab.org/data/).

5.8 Supplementary Materials

1. Supplementary Notes, merged PDF. Contains 4 tables (detailing video metrics, presenting correlations between vFI features with FI score, presenting correlations between vFI features with age, and presenting correlations between manual FI items with age) and 1 figure (showing correlations between video metrics).
2. Supplementary Methods, excel sheet. Details the manual FI scoring (items and score explanations)
3. Supplementary Video 1: Young mouse (8 weeks old). Sample of open field video of a mouse aged 8 weeks.
4. Supplementary Video 2: Old mouse (134 weeks old). Sample of open field video of a mouse aged 134 weeks.
5. Supplementary Video 3: Flexibility metrics. At each frame, 3 points shown in yellow are estimated: base of head (A), mid-back (B), and base of tail (C). At each frame, the distance between points A and C (dAC, shown in red), distance between point B and the midpoint of line AC (dB, shown in green), and the angle formed by the points ABC (aABC, shown in blue) are calculated.
6. Supplementary Video 4: Rearing metrics. Rearing is called when the nose of the mice (marked with a red dot) crosses the perimeter of the open field (yellow line). Rearing is indicated by the presence of the red square in the upper corner of video.

6 Competing Interests

The authors have no competing interest.

7 Acknowledgements

We thank Kumar Lab members, Sean Deats, Tom Sproule, Brian Geuther, and Keith Sheppard for behavioral testing, data processing, and helpful advice. We thank Shock Center and Churchill Lab members Hannah Donato, Gaven Garland, Mackenzie Leland, and Laura Robinson for frailty indexing and coordinating. We thank Taneli Helenius for editing. We thank JAX Information Technology team members Edwardo Zaborowski, Shane Sanders, Rich Brey, David McKenzie, and Jason Macklin for infrastructure support. This work was funded by The Jackson Laboratory Directors Innovation Fund, National Institute of Health DA041668 (NIDA, V.K.), DA048634 (NIDA, V.K.). and Nathan Shock Centers of Excellence in the Basic Biology of Aging AG38070(NIA, G.C.).

References

1. Mitnitski, A., Mogilner, A. & Rockwood, K. Accumulation of Deficits as a Proxy Measure of Aging. *TheScientificWorldJournal* **1,** 323–36 (Sept. 2001).
2. Whitehead, J. C. *et al.* A clinical frailty index in aging mice: comparisons with frailty index data in humans. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* **69,** 621–632 (2014).
3. Rockwood, K., Fox, R. A., Stolee, P., Robertson, D. & Beattie, B. L. Frailty in elderly people: an evolving concept. *CMAJ* **150,** 489–495. ISSN: 0820-3946. eprint: https://www.cmaj.ca/content. https://www.cmaj.ca/content/150/4/489 (1994).
4. Schultz, M. B. *et al.* Age and life expectancy clocks based on machine learning analysis of mouse frailty. *Nature communications* **11,** 1–12 (2020).
5. Searle, S. D., Mitnitski, A., Gahbauer, E. A., Gill, T. M. & Rockwood, K. A standard procedure for creating a frailty index. *BMC geriatrics* **8,** 24 (2008).
6. Kim, S., Myers, L., Wyckoff, J., Cherry, K. E. & Jazwinski, S. M. The frailty index outperforms DNA methylation age and its derivatives as an indicator of biological age. English. *GeroScience* **39,** 83–92 (Jan. 2017).
7. Kojima, G., Iliffe, S. & Walters, K. Frailty index as a predictor of mortality: a systematic review and meta-analysis. *Age and Ageing* **47,** 193–200. ISSN: 0002-0729. eprint: https://academic.oup.com/ageing/article-pdf/47/2/193/24140683/afx162.pdf. https://doi.org/10.1093/ageing/afx162 (Oct. 2017).
8. Parks, R. *et al.* A Procedure for Creating a Frailty Index Based on Deficit Accumulation in Aging Mice. *The journals of gerontology. Series A, Biological sciences and medical sciences* **67,** 217–27 (Mar. 2012).
9. Rockwood, K. *et al.* A frailty index based on deficit accumulation quantifies mortality risk in humans and in mice. *Scientific reports* **7,** 43068 (2017).
10. Kane, A. E., Ayaz, O., Ghimire, A., Feridooni, H. A. & Howlett, S. E. Implementation of the mouse frailty index. *Canadian journal of physiology and pharmacology* **95,** 1149–1155 (2017).
11. Kane, A. E. *et al.* Factors that Impact on Interrater Reliability of the Mouse Clinical Frailty Index. *The Journals of Gerontology: Series A* **70,** 694–695. ISSN: 1079-5006. eprint: https://academic.oup.com/biomedgerontology/article-pdf/70/6/694/10755313/glv032.pdf. https://doi.org/10.1093/gerona/glv032 (Apr. 2015).
12. Feridooni, H. A., Sun, M. H., Rockwood, K. & Howlett, S. E. Reliability of a Frailty Index Based on the Clinical Assessment of Health Deficits in Male C57BL/6J Mice. *The Journals of Gerontology: Series A* **70,** 686–693. ISSN: 1079-5006. eprint: https://academic.oup.com/biomedgerontology/article-pdf/70/6/686/16744417/glu161.pdf. https://doi.org/10.1093/gerona/glu161 (Sept. 2014).
13. Walsh, R. N. & Cummins, R. A. The Open Field Test: a critical review. *Psychological Bulletin* **83,** 482–504 (1976).
14. Crawley, J. N. *Whats Wrong With My Mouse : Behavioral Phenotyping of Transgenic and Knockout Mice* (Wiley, 2007).
15. Ziegler, L., Sturman, O. & Bohacek, J. Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology* **46** (June 2020).
16. Kumar, V. *et al.* Second-generation high-throughput forward genetic screen in mice to isolate subtle behavioral mutants. *Proceedings of the National Academy of Sciences* **108,** 15557–15564. ISSN: 0027-8424. eprint: https://www.pnas.org/content/108/Supplement_3/15557.full.pdf. https://www.pnas.org/content/108/Supplement_3/15557 (2011).
17. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2324 (1998).

18. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60,** 84–90 (2017).
19. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521,** 436–444 (2015).
20. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.
21. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks* **61,** 85–117 (2015).
22. Raghu, M. & Schmidt, E. A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755* (2020).
23. Geuther, B. *et al.* Robust mouse tracking in complex environments using neural networks. *Communications Biology* **2,** 124 (Mar. 2019).
24. Geuther, B. Q. *et al.* Action detection using a neural network elucidates the genetics of mouse grooming behavior. *Elife* **10,** e63207 (2021).
25. Sheppard, K. *et al.* Gait-level analysis of mouse open field behavior using deep learning-based pose estimation. *bioRxiv*. eprint: https://www.biorxiv.org/content/early/2020/12/30/2020.12.29.424780.full.pdf. https://www.biorxiv.org/content/early/2020/12/30/2020.12.29.424780 (2020).
26. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience* **21** (Sept. 2018).
27. Datta, S. QA: Understanding the composition of behavior. *BMC Biology* **17** (Dec. 2019).
28. Baumann, C., Kwak, D. & Thompson, L. Sex-specific components of frailty in C57BL/6 mice. *Aging* **11** (July 2019).
29. Sampathkumar, N. K. *et al.* Widespread sex dimorphism in aging and age-related diseases. *Human genetics* **139,** 333–356 (2020).
30. Austad, S. N. in *Handbook of the Biology of Aging* 479–495 (Elsevier, 2011).
31. Austad, S. N. & Fischer, K. E. Sex differences in lifespan. *Cell metabolism* **23,** 1022–1033 (2016).
32. Sukoff Rizzo, S. J. *et al.* Assessing healthspan and lifespan measures in aging mice: optimization of testing protocols, replicability, and rater reliability. *Current protocols in mouse biology* **8,** e45 (2018).
33. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67,** 301–320 (2005).
34. Breiman, L. Random forests. *Machine learning* **45,** 5–32 (2001).
35. Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20,** 273–297 (1995).
36. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics,* 1189–1232 (2001).
37. Crainiceanu, C. M. & Ruppert, D. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66,** 165–185 (2004).
38. Pappas, L. & Nagy, T. The translation of age-related body composition findings from rodents to humans. *European Journal of Clinical Nutrition* **73** (Oct. 2018).
39. Zhou, Y. *et al.* The detection of age groups by dynamic gait outcomes using machine learning approaches. *Scientific Reports* **10** (Mar. 2020).
40. Skiadopoulos, A., Moore, E. E., Sayles, H. R., Schmid, K. K. & Stergiou, N. Step width variability as a discriminator of age-related gait changes. en. *Journal of NeuroEngineering and Rehabilitation* **17,** 41. ISSN: 1743-0003. https://jneuroengrehab.biomedcentral.com/articles/10.1186/s12984-020-00671-9 (2020) (Dec. 2020).
41. Tarantini, S. *et al.* Age-Related Alterations in Gait Function in Freely Moving Male C57BL/6 Mice: Translational Relevance of Decreased Cadence and Increased Gait Variability. *The Journals of Gerontology: Series A* **74,** 1417–1421 (Nov. 2018).

42. Bair, W.-N. *et al.* Of Aging Mice and Men: Gait Speed Decline Is a Translatable Trait, With Species-Specific Underlying Properties. *The Journals of Gerontology: Series A* **74,** 1413–1416 (Jan. 2019).
43. Zhang, H., Zimmerman, J., Nettleton, D. & Nordman, D. J. Random forest prediction intervals. *The American Statistician,* 1–15 (2019).
44. Meinshausen, N. Quantile regression forests. *Journal of Machine Learning Research* **7,** 983–999 (2006).
45. Athey, S., Tibshirani, J., Wager, S., *et al.* Generalized random forests. *The Annals of Statistics* **47,** 1148–1178 (2019).
46. Friedman, J. H., Popescu, B. E., *et al.* Predictive learning via rule ensembles. *Annals of Applied Statistics* **2,** 916–954 (2008).
47. Apley, D. W. & Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82,** 1059–1086 (2020).
48. Mizrahi-Lehrer, E., Cepeda-Valery, B. & Romero-Corral, A. in *Handbook of Anthropometry: Physical Measures of Human Form in Health and Disease* (ed Preedy, V. R.) 385–395 (Springer New York, New York, NY, 2012). ISBN: 978-1-4419-1788-1. https://doi.org/10.1007/978-1-4419-1788-1_21.
49. Pappas, L. E. & Tim R, N. The translation of age-related body composition findings from rodents to humans. *European journal of clinical nutrition* **73,** 172–178 (Feb. 2019).
50. Huffman, D. M. & Barzilai, N. Role of visceral adipose tissue in aging. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1790,** 1117–1123 (2009).
51. Gerbaix, M., Metz, L., Ringot, E. & Courteix, D. Visceral fat mass determination in rodent: Validation of dual-energy X-ray absorptiometry and anthropometric techniques in fat and lean rats. *Lipids in health and disease* **9,** 140 (Dec. 2010).
52. Imagama, S. *et al.* Back muscle strength and spinal mobility are predictors of quality of life in middle-aged and elderly males. *European Spine Journal* **20,** 954–961 (2011).
53. Holguin, N., Aguilar, R., Harland, R. A., Bomar, B. A. & Silva, M. J. The aging mouse partially models the aging human spine: lumbar and coccygeal disc height, composition, mechanical properties, and Wnt signaling in young and old mice. *Journal of Applied Physiology* **116.** PMID: 24790018, 1551–1560. eprint: https://doi.org/10.1152/japplphysiol.01322.2013. https://doi.org/10.1152/japplphysiol.01322.2013 (2014).
54. Crawley, J. *et al.* Behavioral phenotypes of inbred mouse strains: implications and recommendations for molecular studies. *Psychopharmacology* **132,** 107–24 (Aug. 1997).
55. Singh, P. P., Demmitt, B. A., Nath, R. D. & Brunet, A. The Genetics of Aging: A Vertebrate Perspective. *Cell* **177,** 200–220. ISSN: 0092-8674. http://www.sciencedirect.com/science/article/pii/S0092867419302211 (2019).
56. Pereira, T. D., Shaevitz, J. W. & Murthy, M. Quantifying behavior to understand the brain. *Nature neuroscience,* 1–13 (2020).
57. Mathis, A., Schneider, S., Lauer, J. & Mathis, M. W. A primer on motion capture with deep learning: principles, pitfalls, and perspectives. *Neuron* **108,** 44–65 (2020).
58. Bates, D., Maechler, M. & Bolker, B. Walker., S. Fitting linear mixed-effects models using lme4. *J Stat Softw* **67,** 1–48 (2015).
59. Kenward, M. G. & Roger, J. H. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics,* 983–997 (1997).
60. Fai, A. H.-T. & Cornelius, P. L. Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation* **54,** 363–378 (1996).

8 Supplementary Materials

Table S1: Video Features

Video Metrics			
Category	Name	Description	Units
Open Field	distance_cm	Sum of locomotor activity.	cm
Open Field	center_time_secs	Sum of time spent in center.	sec
Open Field	periphery_time_secs	Sum of time spent along any wall.	sec
Open Field	corner_time_secs	Sum of time spent in any corner.	sec
Open Field	center_distance_cm	Average distance from center across the video.	cm
Open Field	periphery_distance_cm	Average distance from nearest periphery across the video.	cm
Open Field	corner_distance_cm	Average distance from nearest corners across the video.	cm
Open Field	grooming_number_bouts	Sum of all grooming bouts in video.	~
Open Field	grooming_duration_secs	Average length of grooming bouts.	sec
Gait	angular_velocity	The first derivative of angle of a mouse, determined by the vector connecting the mouse's base of tail to its base of neck	deg/second
Gait	lateral_displacement	The difference between the minimum and maximum values of a reference point's perpendicular distance from the mouse's displacement vector for a stride for each frame of a stride, normalized by the mouse's body length. The referece points used are nose, base of tail, and tail tip.	~
Gait	limb_duty_factor	The amount of time that the paw is in contact with the ground divided by the full stride time, calculated and averaged for each hind paw.	~
Gait	speed_cm_per_sec	Speed is determined by the base of tail point.	cm/second
Gait	step_length	The distance that the right hind paw travels past the previous opposite paw strike. Step_length 1 uses left hind paw strike, while step_length2 uses right hind paw strike.	cm
Gait	step_width	The length of the shortest line segment that connects the right hind paw strike to the line that connects the left hind paw's toe-off location to its subsequent foot strike position.	cm
Gait	stride_length	The full distance that the left hind paw travels for a stride, from toe-off to foot-strike.	cm
Gait	temporal_symmetry	The difference in time between the left and right hindpaw strike, divided by the total strike time.	~

Video Features			
Category	Name	Description	Units
Gait	stride_count	Sum of all recorded strides in video.	strides
Gait	distance_cm_sc	Sum of locomotor activity, normalized by time spent in open field.	cm/second
Engineered	dAC	Distance between base of head and base of tail, normalized by the max dAC recorded.	cm
Engineered	dB	The distance between the mid-back point and the midpoint of the line AC.	cm
Engineered	aABC	The angle between the base of head point, mid-back point, and base of tail point.	degrees
Engineered	width	Width of the ellipse fit for the mouse calculated for all frames.	cm
Engineered	length	Length of the ellipse fit for the mouse calculated for all frames.	cm
Engineered	rearpaw	The distance between rearpaws calculated for all frames.	cm
Engineered	rear_count	Sum of rearing bouts in video.	~
Engineered	avg_rear_len	Average length of rearing bouts in video.	sec
End of Table			

Table S2: Feature Correlation with FI score

vFI feauture correlation (Pearson) with FI score						
Features	All Mice		Males		Females	
	Cor	p_{value}	Cor	p_{value}	Cor	p_{value}
median_width	0.636	0.000	0.616	0.000	0.538	0.000
tip_tail_lateral_displacement_iqr	0.628	0.000	0.619	0.000	0.577	0.000
tip_tail_lateral_displacement_stdev	0.620	0.000	0.638	0.000	0.529	0.000
median_step_width	0.620	0.000	0.573	0.000	0.575	0.000
step_width_stdev	0.610	0.000	0.574	0.000	0.582	0.000
median_rearpaw	0.605	0.000	0.546	0.000	0.590	0.000
step_length1_iqr	0.580	0.000	0.642	0.000	0.335	0.000
step_length2_iqr	0.567	0.000	0.621	0.000	0.337	0.000
step_length2_stdev	0.562	0.000	0.605	0.000	0.417	0.000
temporal_symmetry_stdev	0.545	0.000	0.560	0.000	0.420	0.000
step_width_iqr	0.528	0.000	0.496	0.000	0.513	0.000
dB_median	0.519	0.000	0.445	0.000	0.528	0.000
stride_length_stdev	0.517	0.000	0.520	0.000	0.451	0.000
dB_nongait_median	0.512	0.000	0.486	0.000	0.489	0.000
base_tail_lateral_displacement_iqr	0.509	0.000	0.511	0.000	0.472	0.000
step_length1_stdev	0.506	0.000	0.569	0.000	0.251	0.000
median_speed_cm_per_sec	-0.498	0.000	-0.514	0.000	-0.416	0.000
stride_length_iqr	0.486	0.000	0.467	0.000	0.432	0.000
temporal_symmetry_iqr	0.485	0.000	0.499	0.000	0.357	0.000
median_length	0.466	0.000	0.350	0.000	0.489	0.000
limb_duty_factor_stdev	0.444	0.000	0.490	0.000	0.208	0.002
median_stride_length	-0.433	0.000	-0.513	0.000	-0.155	0.019
avg_bout_len	-0.427	0.000	-0.454	0.000	-0.302	0.000
base_tail_lateral_displacement_stdev	0.422	0.000	0.427	0.000	0.387	0.000
speed_cm_per_sec_iqr	-0.372	0.000	-0.399	0.000	-0.347	0.000
dAC_nongait_median	0.352	0.000	0.258	0.000	0.362	0.000
limb_duty_factor_iqr	0.349	0.000	0.370	0.000	0.168	0.011
speed_cm_per_sec_stdev	-0.318	0.000	-0.325	0.000	-0.332	0.000
rears_0_5	-0.315	0.000	-0.303	0.000	-0.269	0.000
dAC_standard deviation	-0.295	0.000	-0.292	0.000	-0.226	0.001
dAC_min	0.294	0.000	0.278	0.000	0.345	0.000
median_tip_tail_lateral_displacement	0.279	0.000	0.207	0.000	0.507	0.000
nose_lateral_displacement_stdev	-0.265	0.000	-0.206	0.000	-0.217	0.001
median_step_length1	-0.263	0.000	-0.329	0.000	-0.020	0.759
aABC_min	0.243	0.000	0.249	0.000	0.276	0.000
aABC_median	-0.241	0.000	-0.189	0.001	-0.192	0.004
nose_lateral_displacement_iqr	-0.221	0.000	-0.159	0.004	-0.164	0.014
rear_count	-0.192	0.000	-0.248	0.000	-0.093	0.161
median_temporal_symmetry	0.191	0.000	0.192	0.001	0.203	0.002
aABC_nongait_standard_deviation	-0.190	0.000	-0.285	0.000	-0.127	0.056
aABC_standard_deviation	-0.184	0.000	-0.261	0.000	-0.176	0.008
rears_5_10	-0.183	0.000	-0.194	0.000	-0.116	0.082
corner_distance_cm	-0.180	0.000	-0.166	0.003	-0.133	0.045
median_step_length2	-0.176	0.000	-0.280	0.000	0.084	0.208
Distance cm/sc	-0.174	0.000	-0.183	0.001	-0.093	0.162

vFI Feature correlation (Pearson) with FI score (Table continued)						
Features	All Mice		Males		Females	
	Cor	p_{value}	Cor	p_{value}	Cor	p_{value}
angular_velocity_iqr	0.157	0.000	0.205	0.000	0.105	0.115
periphery_distance_cm	-0.153	0.000	-0.182	0.001	-0.060	0.370
distance_cm	-0.147	0.001	-0.182	0.001	-0.054	0.421
median_base_tail_lateral_displacement	0.147	0.001	0.096	0.085	0.444	0.000
dAC_nongait_min	0.146	0.001	0.152	0.006	0.211	0.001
aABC_nongait_median	-0.142	0.001	-0.109	0.050	-0.167	0.012
grooming_number_bouts	-0.137	0.001	-0.123	0.028	-0.086	0.196
dB_nongait_min	0.117	0.006	0.136	0.014	0.144	0.030
grooming_duration_secs	-0.109	0.010	-0.075	0.183	-0.133	0.046
corner_time_secs	-0.100	0.019	-0.035	0.529	-0.140	0.035
center_distance_cm	-0.087	0.042	-0.119	0.033	-0.025	0.704
median_limb_duty_factor	-0.086	0.044	-0.088	0.112	-0.068	0.309
median_nose_lateral_displacement	-0.081	0.057	-0.009	0.877	-0.102	0.124
center_time_secs	0.067	0.120	0.047	0.407	0.061	0.358
dB_standard_deviation	0.063	0.140	0.010	0.853	0.061	0.359
dAC_median	0.058	0.177	0.014	0.803	0.194	0.003
dB_nongait_standard_deviation	0.053	0.213	-0.009	0.867	0.150	0.024
periphery_time_secs	-0.053	0.215	-0.038	0.500	-0.067	0.312
aABC_nongait_min	0.047	0.270	0.037	0.510	0.066	0.323
dB_max	0.045	0.294	0.018	0.749	-0.034	0.608
stride_count	-0.035	0.409	0.020	0.713	0.025	0.707
angular_velocity_stdev	0.035	0.414	0.086	0.124	-0.047	0.481
dAC_nongait_standard_deviation	-0.034	0.425	-0.086	0.124	0.027	0.689
dB_nongait_max	0.024	0.579	0.015	0.795	0.026	0.694
median_angular_velocity	-0.004	0.930	-0.039	0.488	0.090	0.177
End of Table						

Table S3: vFI feature correlation (Pearson) with age

Feature Correlation with Age						
Features	All Mice		Males		Females	
	Cor	p_{value}	Cor	p_{value}	Cor	p_{value}
tip_tail_lateral_displacement_iqr	0.764	0.000	0.750	0.000	0.757	0.000
tip_tail_lateral_displacement_stdev	0.737	0.000	0.740	0.000	0.694	0.000
step_width_stdev	0.726	0.000	0.697	0.000	0.727	0.000
median_width	0.697	0.000	0.689	0.000	0.663	0.000
median_step_width	0.697	0.000	0.652	0.000	0.717	0.000
step_length2_stdev	0.678	0.000	0.723	0.000	0.570	0.000
median_rearpaw	0.675	0.000	0.623	0.000	0.702	0.000
step_length2_iqr	0.668	0.000	0.717	0.000	0.499	0.000
step_length1_iqr	0.660	0.000	0.721	0.000	0.461	0.000
step_width_iqr	0.631	0.000	0.594	0.000	0.655	0.000
temporal_symmetry_stdev	0.625	0.000	0.643	0.000	0.523	0.000
stride_length_stdev	0.614	0.000	0.639	0.000	0.549	0.000
base_tail_lateral_displacement_iqr	0.606	0.000	0.576	0.000	0.640	0.000
step_length1_stdev	0.596	0.000	0.655	0.000	0.386	0.000
dB_nongait_median	0.584	0.000	0.577	0.000	0.552	0.000
stride_length_iqr	0.573	0.000	0.567	0.000	0.525	0.000
temporal_symmetry_iqr	0.566	0.000	0.564	0.000	0.490	0.000
median_speed_cm_per_sec	-0.541	0.000	-0.533	0.000	-0.510	0.000
median_length	0.536	0.000	0.444	0.000	0.590	0.000
dB_median	0.525	0.000	0.456	0.000	0.556	0.000
limb_duty_factor_stdev	0.509	0.000	0.546	0.000	0.319	0.000
avg_bout_len	-0.508	0.000	-0.542	0.000	-0.388	0.000
base_tail_lateral_displacement_stdev	0.503	0.000	0.471	0.000	0.543	0.000
median_stride_length	-0.450	0.000	-0.530	0.000	-0.177	0.007
dAC_standard deviation	-0.393	0.000	-0.392	0.000	-0.338	0.000
dAC_nongait_median	0.390	0.000	0.310	0.000	0.418	0.000
limb_duty_factor_iqr	0.383	0.000	0.403	0.000	0.224	0.001
dAC_min	0.380	0.000	0.366	0.000	0.432	0.000
speed_cm_per_sec_iqr	-0.374	0.000	-0.381	0.000	-0.377	0.000
median_tip_tail_lateral_displacement	0.371	0.000	0.252	0.000	0.675	0.000
rears_0_5	-0.364	0.000	-0.326	0.000	-0.376	0.000
speed_cm_per_sec_stdev	-0.314	0.000	-0.320	0.000	-0.327	0.000
aABC_min	0.310	0.000	0.330	0.000	0.317	0.000
nose_lateral_displacement_stdev	-0.297	0.000	-0.279	0.000	-0.202	0.002
aABC_nongait_standard_deviation	-0.267	0.000	-0.315	0.000	-0.265	0.000
aABC_standard_deviation	-0.267	0.000	-0.304	0.000	-0.311	0.000
median_step_length1	-0.262	0.000	-0.321	0.000	-0.035	0.601
nose_lateral_displacement_iqr	-0.254	0.000	-0.237	0.000	-0.146	0.027
median_temporal_symmetry	0.241	0.000	0.248	0.000	0.244	0.000
median_base_tail_lateral_displacement	0.226	0.000	0.119	0.032	0.600	0.000
rears_5_10	-0.210	0.000	-0.195	0.000	-0.193	0.003
grooming_number_bouts	-0.196	0.000	-0.149	0.007	-0.211	0.001
aABC_median	-0.188	0.000	-0.174	0.002	-0.092	0.167
rear_count	-0.188	0.000	-0.224	0.000	-0.113	0.089
grooming_duration_secs	-0.173	0.000	-0.127	0.024	-0.231	0.000

vFI feature correlation (Pearson) with age (Table continued)						
Features	All Mice		Males		Females	
	Cor	p_{value}	Cor	p_{value}	Cor	p_{value}
angular_velocity_iqr	0.160	0.000	0.175	0.002	0.182	0.006
corner_distance_cm	-0.151	0.000	-0.089	0.112	-0.191	0.004
center_distance_cm	-0.147	0.001	-0.188	0.001	-0.074	0.267
distance_cm	-0.147	0.001	-0.154	0.006	-0.098	0.140
median_step_length2	-0.142	0.001	-0.231	0.000	0.110	0.097
Distance cm/sc	-0.133	0.002	-0.157	0.004	-0.025	0.703
dAC_nongait_min	0.132	0.002	0.181	0.001	0.120	0.071
dB_nongait_min	0.131	0.002	0.214	0.000	0.054	0.420
periphery_distance_cm	-0.126	0.003	-0.114	0.042	-0.097	0.146
dAC_nongait_standard_deviation	-0.120	0.005	-0.188	0.001	-0.033	0.625
median_nose_lateral_displacement	-0.112	0.009	-0.083	0.137	-0.061	0.358
median_limb_duty_factor	-0.101	0.018	-0.093	0.095	-0.103	0.123
aABC_nongait_median	-0.091	0.033	-0.093	0.096	-0.065	0.330
dB_standard_deviation	0.083	0.051	0.070	0.205	0.036	0.590
dB_max	0.076	0.074	0.065	0.242	-0.018	0.784
corner_time_secs	-0.066	0.125	0.032	0.563	-0.166	0.012
median_angular_velocity	0.040	0.350	0.001	0.990	0.163	0.014
dB_nongait_standard_deviation	-0.037	0.383	-0.122	0.029	0.098	0.143
aABC_nongait_min	0.035	0.409	0.030	0.586	0.039	0.562
dAC_median	0.034	0.432	-0.023	0.673	0.202	0.002
center_time_secs	-0.028	0.510	-0.075	0.179	0.007	0.919
angular_velocity_stdev	0.023	0.584	0.040	0.476	0.005	0.943
stride_count	-0.023	0.596	0.022	0.697	0.032	0.632
dB_nongait_max	0.014	0.736	-0.007	0.893	0.042	0.527
periphery_time_secs	-0.013	0.764	0.050	0.374	-0.097	0.146
End of Table						

Table S4: Manual FI item correlation with age

Features	All Mice		Males		Females	
	Correlation	p-value	Correlation	p-value	Correlation	p-value
Alopecia	0.137	0.001	0.096	0.084	0.422	0.000
Loss of fur colour	0.451	0.000	0.469	0.000	0.191	0.004
Coat condition	0.633	0.000	0.687	0.000	0.414	0.000
Tremor	0.149	0.000	0.071	0.204	0.287	0.000
Tail stiffening	0.105	0.014	0.108	0.052		
Vision loss (Visual Placing)	0.113	0.008	0.122	0.028	0.229	0.000
Breathing rate/depth	0.241	0.000	0.243	0.000	0.140	0.036
Gait disorders	0.619	0.000	0.614	0.000	0.516	0.000
Body condition	0.454	0.000	0.452	0.000	0.356	0.000
Eye discharge/swelling	0.045	0.287	0.017	0.758	0.096	0.148
Piloerection	0.697	0.000	0.647	0.000	0.719	0.000
Menace reflex	-0.032	0.446	-0.021	0.711	0.045	0.495
Cataracts	-0.036	0.401	-0.058	0.294	0.004	0.953
Malocclusions	0.028	0.508	0.044	0.426	-0.012	0.856
Distended abdomen	0.490	0.000	0.443	0.000	0.496	0.000
Kyphosis	0.658	0.000	0.643	0.000	0.595	0.000
Tumours	0.233	0.000	0.204	0.000	0.249	0.000
Vestibular disturbance	0.286	0.000	0.255	0.000	0.298	0.000
Microphthalmia	-0.007	0.861	0.009	0.877	-0.031	0.644
Righting Reflex	0.255	0.000	0.240	0.000	0.271	0.000
Dermatitis	0.071	0.097	0.098	0.078	0.068	0.305
Loss of whiskers	0.326	0.000	0.396	0.000	0.374	0.000
Corneal opacity	-0.010	0.816	0.007	0.905	0.005	0.944
Nasal discharge						
Rectal prolapse						
Vaginal/uterine/						
Diarrhea						

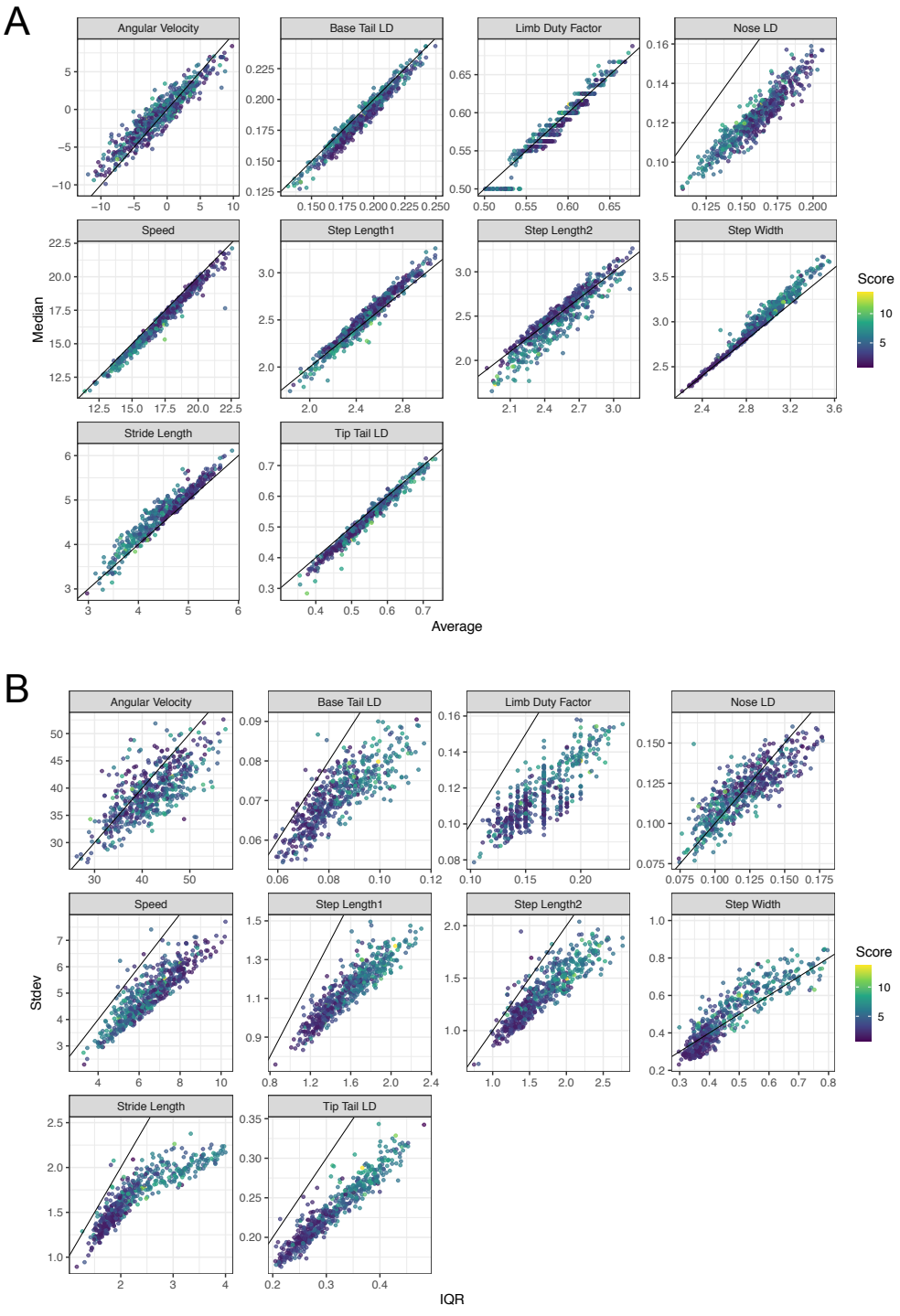

Figure S1: Correlation between video metrics. (A) Correlation between average/mean (x-axis) and median (y-axis) video gait metrics. The diagonal line corresponds to maximum correlation i.e 1. (B) Correlation between inter-quartile range (IQR, x-axis) and standard deviation (Stdev, y-axis) video gait metrics. The diagonal line corresponds to maximum correlation i.e 1. A tight wrap of points around the diagonal line indicates a high correlation between mean and median or IQR and standard deviation for the respective metric.