Original Paper

Machine Learning Ensemble Investigates Age in the Transcriptomic Response to Spaceflight in Murine Mammary Tissue: Observational Study

James A Casaletto[1], BS, MS, PhD; Tyler Zhao[2]; Jay Yeung[2]; Abigail Lee[2]; Amaan Ansari[2,3], BSc; Amber Fry[2]; Arnav Mishra[2]; Ayush Raj[2]; Kathryn Sun[2]; Sofia Lendahl[2], BA; Willy Guan[2]; Melissa S Cline[4], PhD; Sylvain V Costes[5]

[1]Blue Marble Space Institute of Science, Seattle, WA, United States
[2]Student Association for Applied Statistics (SAAS), University of California, Berkeley, Berkeley, CA, United States
[3]University of Mannheim, Mannheim, Germany
[4]Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, United States
[5]NASA Ames, Mountain View, CA, United States

Corresponding Author:

James A Casaletto, BS, MS, PhD
Blue Marble Space Institute of Science
600 1st Ave, First Floor
Seattle, WA 98104
United States
Phone: 1 206 775 8787
Email: james.casaletto@gmail.com

Related Articles:
Preprint (bioRxiv): https://www.biorxiv.org/content/10.1101/2025.02.17.638732v1
Peer-Review Report by Sylvester Sakilay, Mitchell Collier, Arya Rahgozar, Toba Olatoye, Simon Muhindi Savai, Myron Pulier, Randa Salah Gomaa Mahmoud, Clara Amaka Nkpoikanke Akpan, Sayan Mitra, Julie Moonga: https://bio.jmirx.org/2025/1/e75688
Authors' Response to Peer-Review Reports: https://bio.jmirx.org/2026/1/e88583

Abstract

Background: Spaceflight presents unique environmental stressors, such as microgravity and radiation, that significantly affect biological systems at the molecular, cellular, and organismal levels. Astronauts face an increased risk of developing cancer due to exposure to ionizing radiation and other spaceflight-related factors. Age plays a crucial role in the body's response to the cellular stresses that lead to cancer, with younger organisms generally exhibiting more efficient response mechanisms than older ones. The vast majority of research investigating breast cancer risk from spaceflight uses cell lines exposed to simulated radiation and microgravity, but cell lines cannot capture the combinatorial response expressed across tissues, organs, and systems to real radiation and microgravity in space.

Objective: The primary objective of this in silico observational study is to characterize the molecular response to spaceflight of in vivo murine mammary tissue. We use an ensemble of linear binary classifiers to identify the molecular biomarkers enriched in this response using mice flown on the International Space Station. The secondary objective is to determine if age plays a role in this response.

Methods: The National Aeronautics and Space Administration (NASA) Open Science Data Repository has curated transcriptomic data obtained from 10 BALB/cAnNTac female mice flown on the International Space Station and 33 control mice kept on earth (OSD-511). In this observational study focused on two age groups (old/young), we used an ensemble of 4 machine learning binary classifiers with linear decision boundaries (logistic regression, support vector machine, stochastic gradient descent, and single-layer perceptron) to analyze gene expression profiles to predict age (old vs young) and condition (spaceflight vs ground control). Using the genes our ensemble identified as most predictive, we performed pathway enrichment analysis to investigate the molecular pathways involved in spaceflight-related health risks, particularly in the context of breast cancer.

Results: The pathway enrichment analyses revealed age-differentiated responses to spaceflight (false discovery rate–adjusted q values<.05). Among the 10 mice flown in space, younger mice exhibited significantly enriched pathways related to lipid metabolism and inflammatory stress signaling. All space-flown mice demonstrated evidence of adaptation in retinoid metabolism and peroxisome proliferator-activated receptor signaling in response to microgravity and radiation relative to their 33 ground control counterparts.

Conclusions: Spaceflight-induced breast cancer risk manifests through distinct age-specific mechanisms: younger individuals face risk through maladaptive metabolic hyperactivity and oxidative cycling, while older individuals are vulnerable due to impaired stress responses and accumulated metabolic dysfunction. Both age groups ultimately face elevated carcinogenic potential through different but converging pathways. These findings highlight the critical role of age in modulating the response to spaceflight-induced stress and suggest that these molecular pathways may contribute to differential outcomes in tissue homeostasis, metabolic disorders, and breast cancer susceptibility.

JMIRx Bio 2026;4:e73041; doi: 10.2196/73041

Keywords: machine learning; spaceflight; mammary tissue; gene expression; mice; breast cancer; feature importance

Introduction

Spaceflight exposes living organisms to a unique set of environmental challenges, including microgravity [1], radiation [2], and altered gas composition [3], which can significantly impact biological systems at the molecular, cellular, and organismal levels. Several systems have been shown to be impacted in both male and female organisms, including the cardiovascular [4], musculoskeletal [5], immune [6], neurologic [7], hepatic [8], and ophthalmologic [9] systems, to name a few. Although there is currently no evidence of increased gynecological cancer incidence among female astronauts [10], earth-based mouse studies using ionizing radiation, including simulated galactic cosmic radiation, suggest that they may face an increased risk of breast cancer when exposed to space radiation [11]. Exposure to ionizing radiation is well established as a risk factor for breast cancer [12], and both microgravity and simulated microgravity have been shown to enhance the tumorigenic potential of breast cancer cells grown in vitro [13-15]. Furthermore, spaceflight disrupts circadian rhythms, and consequent lower levels of melatonin reduce its efficacy in inhibiting cancer cells [16,17]. Mammary cellular response to spaceflight has been shown to differ with age, as younger organisms typically exhibit more efficient cellular repair and adaptive mechanisms than their older counterparts [18]. Adolescent murine mammary glands exposed to ionizing radiation show increased activation of mammary stem cell and Notch signaling pathways, heightened mammary repopulating activity, and an increased propensity to develop estrogen receptor–negative tumors [19]. A history of ionizing radiation to the chest is a risk factor for breast cancer. The Childhood Cancer Survivor Study indicates that breast cancer risk is highest in young women treated for Hodgkin lymphoma, but it is also increased in those who received moderate-dose chest radiation for other pediatric or young adult cancers [20]. In summary, current research suggests that female astronauts are at a higher risk of developing breast cancer than their terrestrial counterparts, with age being a contributing factor to this increased vulnerability.

The vast majority of research into the risk of breast cancer due to spaceflight has been conducted using simulated radiation and microgravity on either female mice or human breast cell lines. Monti et al [21] found that normal and cancerous breast cell response to microgravity varies drastically, depending on whether the cells are adhered or attached in the organoid model. Kannan et al [22] exposed breast cancer cells to simulated microgravity and compared cells exposed to 10 g and 1 g forces and the respective response in proliferation, cell-cell interaction, and formation of 3D structures, migration, and invasiveness. Although in vitro studies are valuable for mechanistic insights, high-throughput screening, and controlled manipulations, they cannot fully replicate the physiological context of an intact organism. Although simulated microgravity and radiation experimentation on cell lines are much less expensive and resource-intensive approaches than controlled spaceflight experiments, they fail to reproduce the full combinatorial spectrum of the spaceflight environment. Sarkar and Pampaloni [23], in their study of bone marrow remodeling and immune dysfunction in space, note that it remains uncertain how well various microgravity simulation methods replicate the conditions of actual microgravity. They also emphasize that differences in equipment may influence experimental reproducibility, as past studies have frequently produced conflicting results [23].

Bioinformatic approaches have been used to study the effect of spaceflight on health. Many methods in bioinformatics, such as genome-wide association studies and differential gene expression analysis, leverage statistical hypothesis testing as a mechanism to discover new insights. Integrating machine learning (ML) into established bioinformatics and computational biology frameworks has significantly advanced the development of predictive models and analytical tools across molecular evolution, proteomics, systems biology, and disease genomics [24]. ML and artificial intelligence (AI) models are becoming more complex, trained on larger datasets, and run on faster hardware. These trends are accelerating adoption across domains, including bioinformatics. Casaletto et al [25] leveraged an ensemble of ML algorithms to identify genes most predictive of lipid density in murine liver tissue. Building accurate models, particularly with high-dimensional predictors such as gene expression, typically benefits from large sample sizes [24]. To mitigate this, researchers use some form of feature selection—a broad

collection of techniques that reduces the dimensionality of the feature space [26,27]. Filtering methods such as coefficient of variation and feature correlation to a target are examples of feature selection techniques. Traditional ML algorithms such as single-layer perceptrons and logistic regression may be considered weak learners in the context of high-dimensional datasets—but, leveraged together in an ensemble, such weak learners can achieve excellent performance [28].

The use of ML to study spaceflight-induced changes in mammary gene expression can offer valuable insights into the mechanisms of breast cancer development. In this study, we examine the gene expression profiles from a controlled in vivo experiment in which young and old mice were exposed to spaceflight. The mammary glands were dissected and the tissue used for transcriptomic analysis. We are repurposing the data from this study to explore the use of traditional ML methods including random forest, logistic regression, support vector machine, and the single-layer perceptron to determine how murine mammary tissue responds to spaceflight and whether age is a factor. Using the coefficients of simple models such as these to determine feature importance makes this approach very transparent and easy to understand, and combining models into an ensemble makes it a powerful and robust approach.

Methods

In this section, we discuss the data on which this research is based and how we preprocessed it for our ML ensemble. We describe the ensemble of ML algorithms we leveraged, how we derived feature importance from the trained models, and how we combined and filtered the results of the models to form a final set of gene results from our experiments.

Ethical Considerations

We used OSD-511 as the source of data for our observational study. All National Aeronautics and Space Administration (NASA) rodent research missions, including Rodent Research Reference Mission 1 (RRRM-1) from whence our data are derived, are required by US federal law to follow strict humane care and use of laboratory animals under the provisions of the Health Research Extension Act of 1985 [29]. As an observational study, our research was conducted on data from an already-published experiment. The authors believe the repurposing of existing datasets not only maximizes the cost-effectiveness of those studies, it also eliminates the need to further expose animals to the conditions of spaceflight and ultimately sacrifice animals for novel research.

Data

In the RRRM-1, a total of 43 female BALB/cAnNTac mice were included in the study, consisting of 21 younger mice (aged 9-12 weeks, YNG) and 22 older mice (aged 32 weeks, OLD). Among the younger mice, 5 were flown in space, 8 were kept in the Animal Enclosure Module (AEM), and 8 were housed in regular vivarium cages (VIV). Vivarium controls are included in spaceflight studies to distinguish the effect of the cage used in spaceflight (ie, AEM) from the ambient effects of spaceflight (eg, radiation, microgravity). In this research, we do not explore that distinction, so we combined the VIV and AEM control groups into a single ground control group called “GC.” For the older mice (OLD), 5 were flown in space, 7 were housed in flight hardware, and 10 in vivarium cages. Note that there was no basal group included in the design of their experiment. After 40 days in space, the mice were safely returned to Earth, given 2 days to recover (Live Animal Return), and then euthanized. Mice flown in space and kept in standard cages are denoted FLT. Table 1 summarizes the distribution of mice in the experiment.

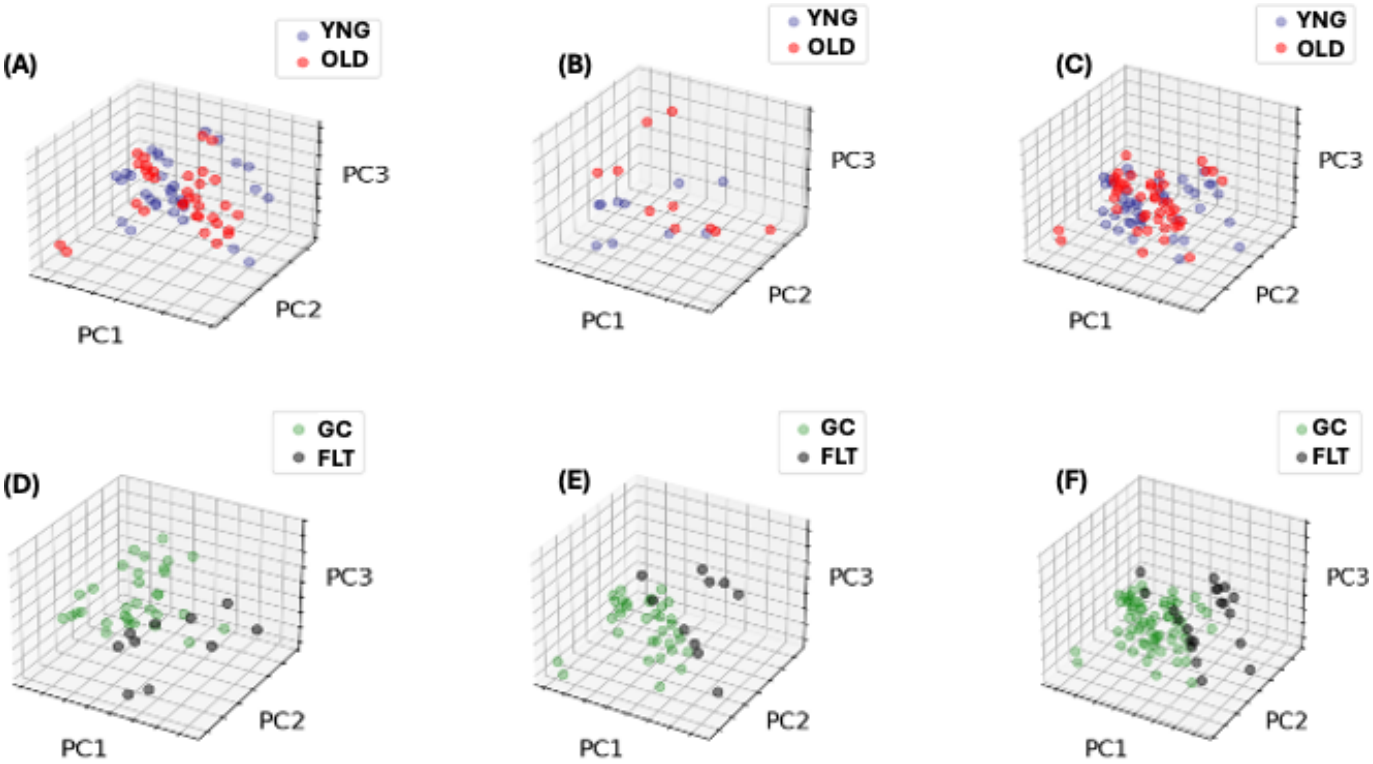
Table 1. Distribution of mice in different experimental groups, including flight habitat (AEM) and vivarium (VIV), which together constitute the overall ground control (GC=AEM+VIV), and spaceflight (FLT) groups for both the old (OLD) and young (YNG) cohorts. Marginal totals are provided in the last column of the table.

	OLD (32 weeks)	YNG (9-12 weeks)	Total
AEM (Animal Enclosure Module)	7	8	15
VIV (vivarium)	10	8	18
GC (ground control)	17	16	33
FLT (spaceflight)	5	5	10

The dataset contains ribo-depleted total RNA sequencing (RNA-seq) data from mammary glands. The sequences for each mouse were aligned once using *Mus musculus* Spliced Transcripts Alignment to a Reference (STAR; version 2.7.10a) and once with RNA-Seq by Expectation-Maximization (RSEM version 1.3.1) to the Ensembl release 107, genome version GRCm39. These data are available in the Open Science Data Repository [30] as dataset OSD-511 [31]. Both datasets (RSEM, STAR) are published with OSD-511.

The principal component analysis (PCA) plots of the data are shown in Figure 1. PCA projections that display approximately linearly separable classes suggest that binary classifiers with linear decision boundaries, such as those in our ensemble, may achieve strong classification performance.

Figure 1. PCA plots for each of the experiments (augmented datasets with RSEM, STAR). Figures 1A-C are PCA plots of the ground control mice, spaceflight mice, and all mice, respectively, and are colorized by age. Figures 1D-F are PCA plots of the young mice, old mice, and all mice, respectively, and are colorized by condition. PC: principal component; PCA: principal component analysis; RSEM: RNA-Seq by Expectation-Maximization; STAR: Spliced Transcripts Alignment to a Reference.



Based on the 3D PCA plot in Figure 1A, age among ground control mice did not seem to be predictable from gene expression with a linear decision boundary. This supportd the use of the control group and provided a neutral baseline for later comparisons. In Figure 1B, gene expression differed between young and old mice in response to spaceflight. Figures 1D and E showed a clear distinction between ground control and spaceflight among young and old mice, respectively. This pattern suggested an age-related response to spaceflight and motivated us to investigate further. Figure 1C did not clearly distinguish young from old mice, but Figure 1F showed a clear separation between the unmarginalized age groups. This suggests that the impact of age on gene expression is not as significant as the impact of spaceflight.

ML model performance generally improves with more data points. Additionally, training and testing must be performed on a sufficient number of data points to accurately quantify model performance. Data augmentation is a collection of methods used to increase the size of a dataset for training and testing. In our research, we combined the RSEM and STAR datasets by creating 2 data points per biological sample: one for the RSEM quantification and one for the STAR quantification. This increased the size of our dataset by a factor of 2, with the caveat that the augmented samples are not independent (see points in Figure 1). Because ML model performance improves with fewer dimensions, we performed the filtering methods described in Table 2 to reduce the dimensionality of the dataset. We removed genes that have nonnumeric values or not-a-number values, genes that do not code for proteins, genes with counts below 30 in 80% of the samples, genes with a coefficient of variation lower than 0.4, and nondifferentially expressed genes at an α level of 0.1.

Table 2. Data-filtering methods applied to this dataset include removing genes with not-a-number values, noncoding genes, and genes that are not correlated to the binary targets (old vs young or ground control vs spaceflight). Columns include the total count of genes before the filter was applied, the total number of genes removed by the filter, and the count of genes remaining after the filter was applied. These filters were executed in order from top to bottom, leaving a total of 750 genes for training our models.

Filtering method	Count before filter	Number removed by filter	Count after filter
Remove genes with not-a-number values	56,840	0	56,840
Remove non–protein-coding genes	56,840	35,159	21,681
Remove noncorrelated genes	21,681	20,931	750

After reducing the dimensionality of the data, we applied three transformations. First, we transformed the data into transcripts per million to account for sequencing depth and gene length, thus making the gene expression values comparable within a sample. Second, we applied a log transformation to stabilize the variance inherent in transcriptomic count data. Third, since coefficient-based ML algorithms require all the feature values to be on the same

scale, we used the StandardScaler method from *scikit-learn* to convert all feature values to z-scores.

Figure 2 shows the graphical summary of the methods we used in our in silico experiments to create sets of genes that are predictive of their respective targets. We introduce the notation "GROUP:target" to denote the experiment where GROUP represents the subsets ground control (GC) and spaceflight (FLT) or the subsets young mice (YNG) and old mice (OLD), and the target represents the binary class age or condition (cnd) that the ML model is trained to predict.

Figure 2. Graphical summary of the methods used in this research. (A) The OSD-511 dataset contains RNA-seq data for mouse mammary tissue. (B) The data were filtered to reduce dimensionality, normalized, log-transformed, and standardized. (C) Data were divided into GC and FLT groups to predict age and divided into YNG and OLD groups to predict condition. (D) Each subset of data was used to build 4 models in the ensemble. (E) Each model generated two sets of genes most predictive of the target. (F) The two sets from each model were unioned into a single set per model. (G) The four sets from each model were majority-intersected to yield the intermediate set of genes per experiment. cnd: condition; FLT: spaceflight; GC: ground control; OLD: old; YNG: young.

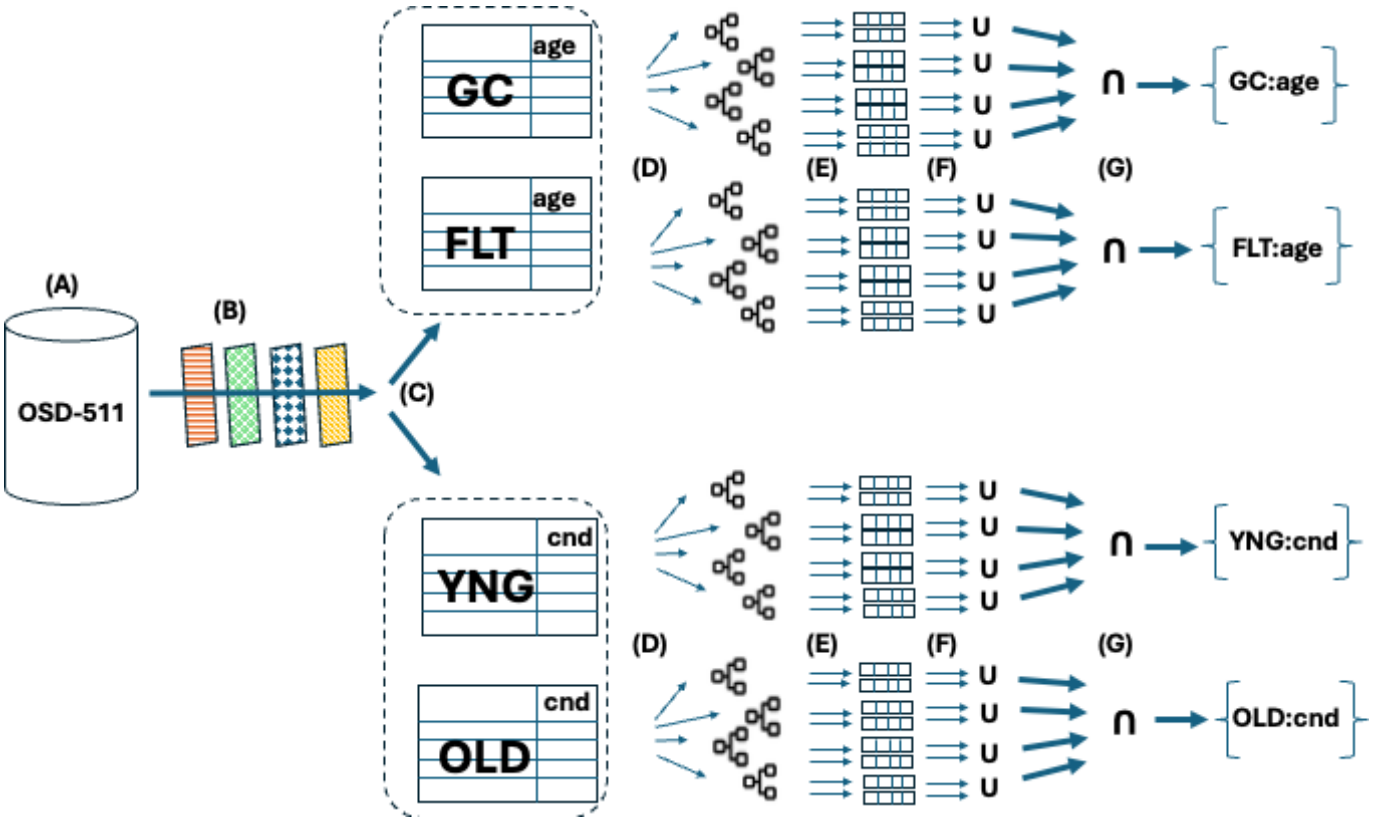
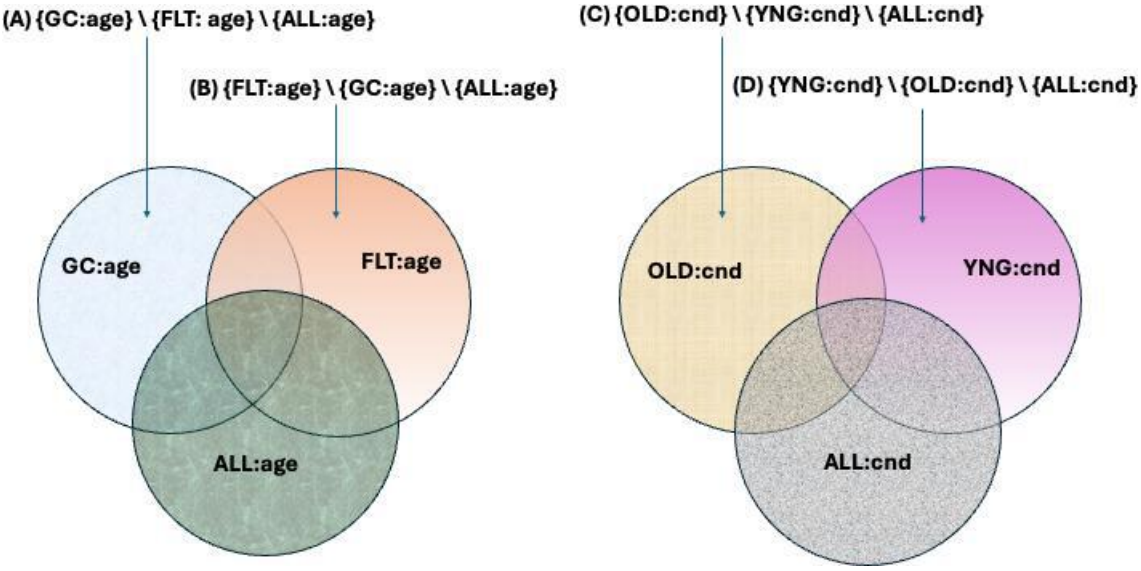


Figure 2 shows all the steps in the pipeline to produce the intermediate result set of genes, which were further processed as described in Figure 3.

Figure 3. Venn diagrams depicting set difference operations to identify genes uniquely predictive of age (A and B) and condition (C and D) for a given subset of mice. In Figure 3A, we remove ALL:age genes and FLT:age genes that intersect with GC:age to obtain those genes that uniquely predict age for ground control mice. These genes are represented by the light blue part of the Venn diagram. Similarly, we remove ALL:age genes and GC:age genes that intersect with FLT:age genes to obtain those genes that uniquely predict age for space-flown mice. These genes are represented by the light orange part of that Venn diagram. In Figures 3C and 3D, we use the same logic to obtain those genes that uniquely predict the condition of old mice in the yellow, textured part of the Venn diagram and those genes that uniquely predict the condition for young mice in the pink part of the Venn diagram. These set operations yielded the final gene results we discuss in the next section. cnd: condition; FLT: spaceflight; GC: ground control; OLD: old; YNG: young.



Algorithms

We leveraged 4 supervised ML algorithms on the gene expression data to predict labels associated with each sample. These models were trained and tested to classify binary labels (spaceflight vs ground control and old vs young) and include stochastic gradient descent (SGD), logistic regression (LR), single-layer perceptron (SLP), and support vector machine (SVM). These models were specifically selected to capture linear decision boundary classification patterns.

The SGD classifier from *scikit-learn* trains a linear classifier using stochastic gradient descent to update the coefficients of the input features. SGD iteratively updates the coefficients based on the gradient of the loss function, using one training sample at a time to compute each gradient step, rather than the whole dataset. We used the *scikit-learn* implementation of SGDClassifier with all default hyperparameters. LR, despite its name, is a binary classification algorithm that provides a probability for the binary target prediction based on a set of discrete or continuous features [32]. Because it does use regression, there are model coefficients associated with the features that may be used for feature importance. We used the *scikit-learn* implementation of LR as a binary classifier with all default values for the hyperparameters. The SLP was developed in the 1950s by Frank Rosenblatt and is the most basic form of neural network [33]. The input features are weighted in a linear combination that can either be sent through a sigmoidal activation function for binary classification or through a linear activation function for regression. Feature importance is conveniently derived directly from the feature weights, which makes the SLP an easy-to-interpret ML algorithm. We used the *scikit-learn* implementation of SLP as a binary classifier with all default hyperparameter values. The SVM was created by Hava Siegelmann and Vladimir Vapnik as a margin-based classifier using so-called support vectors to separate classes in the feature space [34]. Feature importance is derived directly from the coefficients of the support vectors of linear kernels. We used the *scikit-learn* implementation of the linear SVM with all default hyperparameter values.

All four models were trained using a train/test split of 80/20 with GroupShuffleSplit() from *sklearn.model_selection*. This method allows users to specify which samples must be grouped together after the split, permitting us to keep the RSEM and STAR replicates in the same train and test groups and thereby prevent target leakage. The models were validated using the *scikit-learn* implementation of k-fold cross validation, and we used k=5 as the number of folds because we had such few samples. Because of the small number of samples, we repeated the experiments several times using different seeds for the random number generators used throughout the pipeline. We deployed the four classification algorithms as binary classifiers in two experiments: predicting age (OLD vs YNG) and predicting condition (FLT vs GC) using gene expression data as predictors. After training each model, we identified the features most predictive of the classes using the two methods described in the next section.

Per-Model Feature Importance

In our method, we combined multiple ML algorithms into an ensemble classifier to predict either experimental condition (ground control vs spaceflight) or age (young vs old). We quantified feature importance by coefficient magnitude in two parts of the pipeline: cross-validation and a standard train-test split. In the cross-validation setting, the data were partitioned

into 5 folds, and models were then trained on a single fold and evaluated on the other 4 folds. This procedure yielded 5 fitted estimators. For each estimator, *scikit-learn* provided coefficients from which we derived feature importances. We then averaged the importances for each feature across the folds, ranked the features according to this mean value, and kept the top 50 highest-coefficient features. In the train-test approach, we fitted the model to the training set, ranked the coefficients by magnitude and selected the top 50 as the most predictive features. We combined these two gene sets together into a single set of genes using the union set operation and then removed genes overlapping with other experiments as described in the next section.

Per-Experiment Ensemble Voting

Ensemble predictions are commonly aggregated by majority voting [35]. For each experiment, we first formed, for each algorithm, the union of the two feature importance lists. We then applied majority voting across the 4 algorithm-specific unions, retaining genes that were present in at least 3 of them. We obtained the final label predictive set with a difference operation, as described in the next section.

Final Gene Set Formulation

To determine the genes that are most predictive of a target (age or condition) for a given subset of mice (eg, YNG vs OLD or FLT vs GC), we removed those genes that are generally predictive of the target, regardless of their subset. In this way, we identified the marginal set of genes that are uniquely predictive of the target within that subset. For example, in the experiment in which we predicted age, we ran 3 experiments: one in which we used only ground control samples to predict age (GC:age), one in which we used only spaceflight samples to predict age (FLT:age), and one in which we used all the samples combined to predict age (ALL:age). Each of these 3 experiments produced a set of gene results as previously described. In Figure 3, we showed how we formulated our final set of gene results for analysis. We adopted the notation {X} \ {Y} to represent the difference in set membership between sets X and set Y.

Results

In this section, we discuss the final results of our 4 experiments: predicting age for ground control samples, predicting age for spaceflight samples, predicting condition for old samples, and predicting condition for young samples.

Model Performance

Since our models do not classify outcomes as "positive" and "negative" with different associated costs, metrics such as the false positive rate and false negative rate offer limited insight. Given the imbalanced class distribution between ground control and spaceflight groups, accuracy is an inadequate performance measure. To evaluate model performance using a single comprehensive metric, we selected the F_1-score, which represents the harmonic mean of precision and recall, as our primary performance indicator. Table 3 displays the F_1-score (averaged over 5 different random number generator seeds) of each of the 4 classification models in the ensemble for the experiments predicting age in FLT, GC, and ALL groups. The train and test scores were obtained using the 80/20 train/test split data sets, and the cross-validate score is the mean score across the 5 folds.

Table 3. Average F_1-score for training, testing, and cross-validation of each of the classification models (stochastic gradient descent, support vector machine, logistic regression, and single-layer perceptron) for the experiments predicting age (FLT[a]:age, GC[b]:age) for those mice in the FLT and GC groups.

Model and experiment	Train	Test	Cross-validate
Stochastic gradient descent			
FLT:age	1.0	0.96	0.89
GC:age	1.0	0.99	0.98
Support vector machine			
FLT:age	1.0	1.0	0.91
GC:age	1.0	1.0	0.99
Logistic regression			
FLT:age	1.0	1.0	1.0
GC:age	1.0	1.0	1.0
Single-layer perceptron			
FLT:age	1.0	1.0	1.0
GC:age	1.0	0.99	1.0

[a]FLT: spaceflight.
[b]GC: ground control.

Table 4 displays the performance of each of the 4 classification models in the ensemble for the experiments predicting the condition for YNG and OLD groups.

Table 4. Average F_1-score for training, testing, and cross-validation of each of the classification models for the experiments predicting condition (OLD[a]:cnd[b], YNG[c]:cnd) for those mice in the OLD and YNG groups.

Model and experiment	Train	Test	Cross-validate
Stochastic gradient descent			
OLD:cnd	1.0	0.97	0.84
YNG:cnd	1.0	0.99	0.91
Support vector machine			
OLD:cnd	1.0	1.0	0.90
YNG:cnd	1.0	1.0	0.92
Logistic regression			
OLD:cnd	1.0	1.0	1.0
YNG:cnd	1.0	1.0	1.0
Single-layer perceptron			
OLD:cnd	1.0	0.96	0.94
YNG:cnd	1.0	1.0	1.0

[a]OLD: old.
[b]cnd: condition.
[c]YNG: young.

As shown in Tables 3 and 4, all the train scores had a perfect F_1-score, and all but 3 of the test scores in each table were also perfect. The cross-validate score is useful in determining to what extent there is bias in the model due to how the train and test data were split, or how much the model is otherwise overfit. The experiments predicting condition for young mice outperformed the same experiments for old mice. The SLP and LR models outperformed SGD and SVM in all experiments. In Table 2, SGD scored the lowest F_1-scores in both experiments (OLD:cnd, YNG:cnd) predicting the condition. Because we used the majority consensus for our feature voting algorithm, we acknowledge SGD as the weakest learner for those experiments and accept the results from the rest (majority) of the ensemble. After training each model, we identified those genes most predictive of their respective target. We present these results in the next section.

Most Predictive Genes

In this section, we discuss the genes most predictive of the targets for each experiment. Textbox 1 lists the genes most predictive of the label for each of the experiments.

Textbox 1. List of genes most predictive of the target (age, condition [cnd]) for the given subset (GC [ground control], FLT [spaceflight], YNG [young], OLD [old]).

GC:age
- *Aip, Aldh2, Ceacam10, Ciao2b, Clec4d, Csn1s2a, Ctsz, Dmbt1, Gng2, Gstm1, Klk4, Lrrc30, Mrgprb1, Myh8, Nudt9, Or13a27, Pam, Park7, Prom1, Psmc4, Psmd2, Slc5a5, Smyd2, Syngr2, Tle5, Vmn1r38, Wap, Wdr18, Yif1b, Znhit2*

FLT:age
- *Acsm1, Acss2, Adamdec1, Adcy10, Ahsg, Aldoa, Aldob, Ap2b1, Apoa1, Apoa2, Apoa4, Atp1a3, Atp6ap1, Bmp2k, Ces1g, Chrna5, Cps1, Cyp2c29, Cyp2c50, Elovl3, Epyc, Fabp1, Fga, Fgb, Fmo3, Gbp11, Gc, Gnl1, Hadha, Hspa5, Immt, Lmod2, Lrrc59, Maob, Mat1a, Mogat2, Mrpl30, Mtch1, Ncan, Psmb7, Ptk7, Rad23b, Ramp2, Rdh11, Scgb1c1, Serpinf2, Slc10a1, Slc25a3, Slc25a39, Slc27a5, Slc38a3, Ssx2ip, Stfa3, Sult3a1, Tat, Tdrd9, Tmem259, Ugt2b34, Uox, Urod, Zfp747*

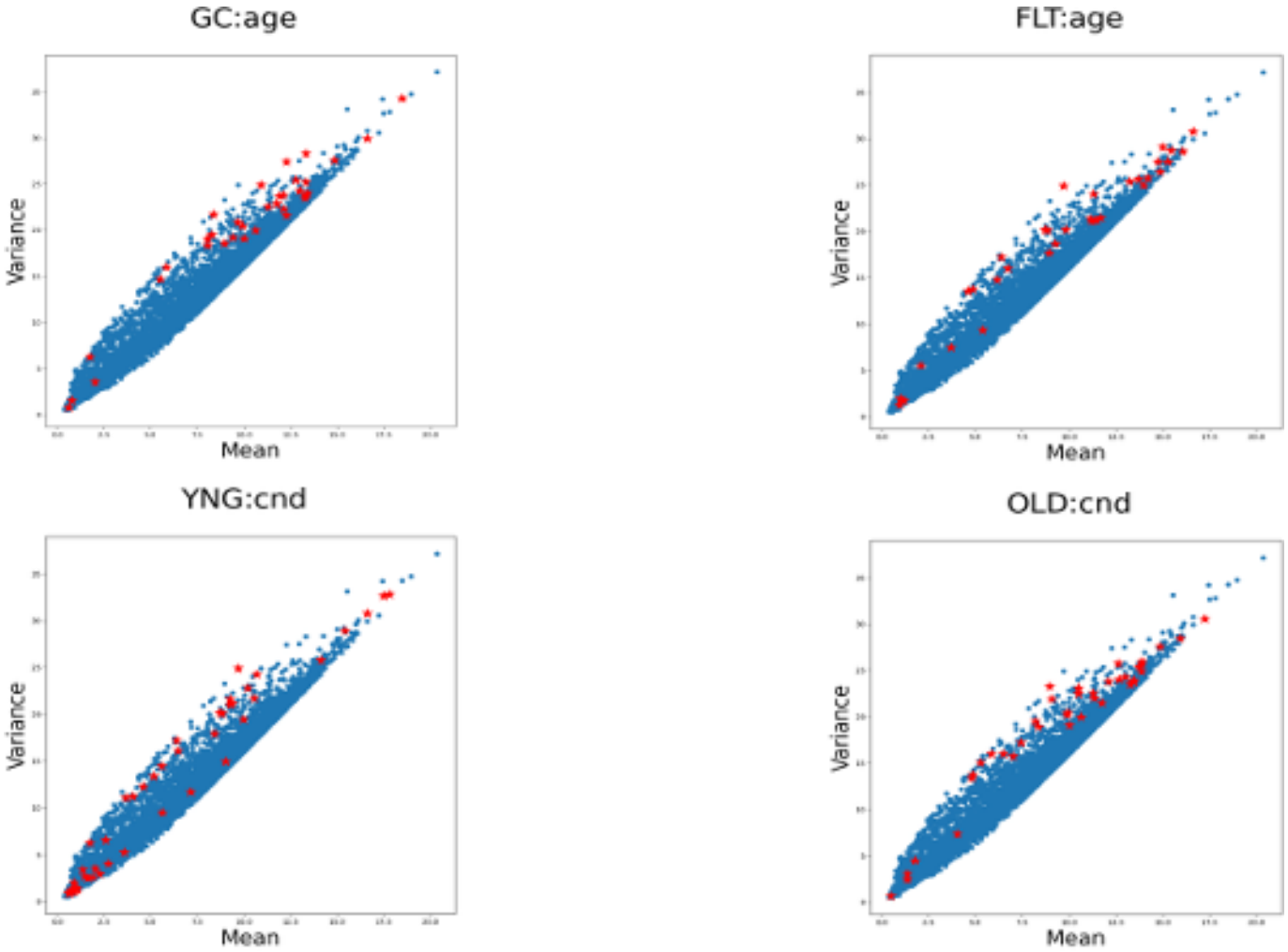
YNG:cnd
- *Aar2, Abcc6, Acot11, Apcdd1, Aspg, Cdcp3, Elovl3, Ergic1, Gale, H1f0, Hspb8, Kcng4, Ltc4s, Maff, Map3k4, Mogat2, Mrpl47, Mrps18a, Ncan, Odad4, Pnpla5, Postn, Ppcs, Prune2, Rdh11, Scd2, Sfxn5, Smtnl2, Tekt1, Tmprss11a, Vstm2b*

OLD:cnd
- *6430571L13Rik, Acad10, Actl6b, Agtr1a, Ambp, B3gnt7, Begain, Calca, Ceacam20, Cuta, Dgat2, Fgf21, Glud1, Igfbp4, Igsf21, Jmjd8, Krt12, Krtap6-7, Map6d1, Mrpl42, Or2y1e, Or51r1, Or56b35, Rgs16, S100a9, Tcap, Trim9, Ttr, Vmn1r32*

The genes listed in constitute the final results of our ML ensemble that resulted from the set operations portrayed in Figure 3.

In Figure 4, we show the distribution of gene expression for the most predictive genes of each experiment across the distribution of all the genes that were used to train the models.

Figure 4. Scatter plots of variance versus mean for the experiments predicting age (top row) and predicting condition (bottom row). The blue points are the background genes (ie, all 750 genes that were used to train the model), and the red points are most predictive of their respective target. cnd: condition; FLT: spaceflight; GC: ground control; OLD: old; YNG: young.



As shown in Figure 4, the distribution of the genes identified by our ML ensemble across the spectrum of expression is approximately uniform. From that, we can infer that the ML algorithms do not portray any bias based on the magnitude (mean or variance) of the distributions of gene counts. This indicates that the models and their ensemble are not vulnerable to the heteroskedastic nature of gene expression count data. Note that the distribution of genes predicting age is different than the distribution of genes predicting condition because we used the 750 genes most correlated to the respective target. We next show which biological pathways are enriched by the GC:age, FLT: age, YNG:cnd, and OLD:cnd gene sets.

Pathway Enrichment Analysis

We submitted our lists of most predictive genes to ShinyGO (version 0.81)—an online pathway enrichment analysis tool [36]—using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [37], a false discovery rate cutoff of .05, and minimum gene set intersection size of 2, and displayed the top 5 most enriched pathways. The results of these analyses are captured in Table 5.

Table 5. Pathway enrichment analyses for the machine learning experiments. All corresponding false discovery rate *q* values were statistically significant to an α level of less than .05.

Experiment and pathways	Genes	False discovery rate *q* value
GC[a]:age		
No enrichment	—[b]	—
FLT[c]:age		
Metabolic pathways	*Ugt2b34, Cyp2c50, Maob Aldoa, Acsm1, Mat1a Elovl3, Cyp2c29, Fmo3*	1.456e-07

Experiment and pathways	Genes	False discovery rate *q* value
	Rdh11, *Uox*, *Urod*, *Cps1* *Aldob*, *Mogat2*, *Tat* *Slc27a5*, *Adcy10*, *Atp6ap1* *Acss2*, *Hadha*	
Fat digestion and absorption	*Apoa1*, *Apoa4*, *Fabp1*, *Mogat2*	0.00037
Biosynthesis of amino acids	*Aldoa*, *Mat1a*, *Cps1*, *Aldob*	0.00264
Peroxisome proliferator-activated receptor signaling pathway	*Apoa1*, *Apoa2*, *Fabp1*, *Slc27a5*	0.00331
Retinol metabolism	*Ugt2b34*, *Cyp2c50* *Cyp2c29*, *Rdh11*	0.00347
YNG[d]:cnd[e]		
Biosynthesis of unsaturated fatty acids	*Elovl3*, *Scd2*	0.02312
Fatty acid metabolism	*Elovl3*, *Scd2*	0.03802
Metabolic pathways	*Ppcs*, *Elovl3*, *Ltc4s*, *Rdh11*, *Scd2*, *Mogat2*, *Gale*	0.00779
	Rdh11, *Scd2*, *Mogat2* *Gale*	
OLD[f]:cnd		
No enrichment	—	—

[a]GC: ground control.
[b]Not applicable.
[c]FLT: spaceflight.
[d]YNG: young.
[e]cnd: condition.
[f]OLD: old.

The most important genes predicting the age in the ground control group (GC:age) and those predicting the condition in the old group (OLD:cnd) did not significantly enrich any of the KEGG pathways. The genes most predictive of age in the spaceflight group (FLT:age) enriched several KEGG pathways, the top 5 of which are shown in . The metabolic pathways enrichment represents a very broad class of biological functions including lipid metabolism, energy metabolism, and xenobiotic metabolism. The peroxisome proliferator-activated receptor (PPAR) signaling pathway represents fatty acid oxidation, lipoprotein metabolism, and an anti-inflammatory response. Because retinoids are antioxidants, the retinol metabolism pathway is likely responding to oxidative stress. The genes most predictive of condition for the young mice (YNG:cnd) also primarily enriched membrane lipid metabolism, inflammatory stress signaling, and overall metabolic capacity. All these pathways being enriched suggests that spaceflight amplifies age-related differences in metabolic flexibility, especially in pathways that manage lipid metabolism in response to inflammation and oxidative stress. In the Discussion section, we will explore this theme further in the context of breast cancer.

Discussion

Principal Findings

In this study, we used a novel approach combining results from an ensemble of 4 linear classifier ML models to predict condition (spaceflight or ground control) and age (young or old) using features derived from gene expression data. The results reveal distinct gene expression signatures that differentiate both age and exposure to spaceflight in mice, revealing some of the molecular mechanisms that may underpin the effects of spaceflight and aging and their potential impact on breast cancer. In this section, we discuss the principal findings of our research in the context of breast cancer, compare our approach to other ML approaches on transcriptomic data, describe strengths and limitations to our methods, and conclude with considerations toward future directions of this research.

Our research finds that the younger mouse cohort mounted a differential response to spaceflight with respect to their older counterparts. One reason for this may be that younger cells have higher plasticity, and therefore their tissue has greater capacity to respond to the environment [38]. Older cells may have blunted responses because they have exhausted their capacity to respond due to accumulated stress [39]. Another reason may be signal saturation: older tissue has chronic low-grade inflammation and is already expressing a stress response to oxidative damage at a baseline [40]. In the context of breast cancer risk due to spaceflight, our research paradoxically suggests that the younger cohort may have an increased risk due to the simultaneous modulation of PPAR signaling and fatty acid biosynthesis. The younger cohort gene expression enriched unsaturated fatty acid metabolism pathways in the *Elovl3* and *Scd2* genes. Galactic cosmic radiation generates reactive oxygen species, which attack unsaturated fatty acids in membranes, leading to lipid peroxidation [41]. Damaged lipids, if left unchecked,

can cause mitochondrial and nuclear membrane damage, leaving cells struggling to maintain basic homeostasis [42]. The genes enriching the PPAR signaling pathway (*Apoa1*, *Apoa2*, *Fabp1*, and *Slc27a5*) are all PPAR-α genes, which promote the breakdown of damaged fatty acids so they may be used as an energy source [43]. This can lead to a vicious cycle whereby fatty acids are synthesized and then oxidized, inducing reactive oxygen species production, which causes more lipid peroxidation [44]. The subsequent proliferation of peroxisomes would put these younger mammary cells under chronic oxidative stress and increase carcinogenic potential [45].

Our research suggests that older mice may be at increased risk of breast cancer for different reasons. In the experiment predicting condition (spaceflight vs ground control) for all mice, their most predictive genes enriched pathways in retinol metabolism and PPAR signaling. The genes enriching the retinol metabolic pathway include *Ugt2b34*, *Cyp2c50*, *Cyp2c29*, and *Rdh11*. The *Rdh11* enzyme, or retinol dehydrogenase 11, synthesizes retinoids, which regulate cell proliferation, promote cell differentiation, and induce apoptosis—all of which help prevent and suppress mammary gland tumor formation [46]. However, the *Cyp2c50* and *Cyp2c29* genes are degradation enzymes in this pathway and lead to retinoid depletion. Moreover, the *Ugt2b34* gene is an excretion enzyme that eliminates active retinoids. The overall metabolic impact on this pathway may lead to the degradation of retinoids, which would greatly increase the risk of developing breast cancer [47]. The simultaneous disruption of PPAR signaling and retinoid metabolism in mammary tissue following spaceflight represents a synergistic increase in breast cancer risk [48-50]. This two-hit disruption is normally more severe in older animals due to depleted antioxidant reserves and reduced metabolic flexibility [51], suggesting that older individuals may face substantially elevated breast cancer risk from spaceflight exposure.

Comparison to Prior Work

Zhang et al [52] built an ML model that leverages a transformer architecture, incorporating phenotype prediction, biomarker discovery, and identification of implicated biological processes into a single model using transcriptomic data as features. Our research provides similar types of analyses, but we use binary classification models for phenotype prediction and two forms of feature importance to identify biomarkers; we also leverage an existing, well-used framework (ie, KEGG pathways) for identifying biological processes. Smith et al [53] use a similar set of data processing steps in their pipeline (converting gene counts to transcripts per million, applying log transformations) in an ML ensemble, but they use regression rather than classification to predict phenotypes. Arnold et al [18] examined the same dataset (OSD-511) as the one explored in this research but used differential gene expression analysis to identify the biomarker genes that distinguish young from old and spaceflight from ground control mice. Differential gene expression analysis is a commonly used technique for high-dimensional data but suffers from multiple test burden and an inability to distinguish between true and spurious correlations.

Strengths and Limitations

The first strength and motivating factor for studying this data is to maximize the utility of underpublished in vivo research in controlled spaceflight experiments. Murine experimentation in space is very costly, time-consuming, and requires sacrificing animals. As an observational study, we obtained real-world insights without further cost and sacrifice. The second strength of our approach is model interpretability. Particularly in the context of predicting biomedical outcomes, using whitebox, linear decision-boundary models such as SGD classifier, SVM, LR, and SLP enables transparency, engenders trust, and provides more straightforward biological insight into a high-dimensional feature space such as gene expression data. The third strength of our approach is the use of simple set operations (union, intersection, and difference) to improve interpretability. The fourth strength of our approach is the use of the KEGG database as a trusted, well-known pathway enrichment analysis database to further promote simplicity and trust.

The first limitation of our approach is that we excluded many ML methods, such as multilayer perceptrons and other deep learning architectures, that may outperform the ones we used at the expense of simplicity and interpretability. The second limitation of our study is the sensitivity of the results to our preprocessing. For example, removing genes that have low counts and are not correlated to the target reduces the signal-to-noise ratio in a high-dimensional feature space. However, because some biological processes are sensitive to slight variations in gene expression, we may have removed some of the genes that contribute to the phenotypes that our models predicted. Filtering out genes that do not code for proteins allows our pathway enrichment analysis to focus on well-understood genes, though again, we understand that noncoding genes may also have contributed to the phenotypes. The third limitation of our study is the paucity of data. We would feel more confident in our results if we could explore a larger and more varied collection of samples. The fourth limitation of our study is the lack of an in vivo or in vitro validation of our findings. Although the gold standard in biomarker identification is the randomized controlled trial, our observational research serves to inform such a study and can restrict the search space of an otherwise very resource-intensive endeavor. The last limitation of our research is that it relies on a single point-in-time snapshot of the mammary transcriptome via bulk RNA-seq. A better approach would be a longitudinal investigation that elucidates time as a contributing factor to spaceflight response.

Future Directions

Our research has identified putative genes and pathways implicated in age-differentiated pathological responses to spaceflight in mammary tissue. Future work may include single-cell RNA sequencing and proteomic sequencing to give higher resolution and downstream validation, respectively. Combining multiple datasets from similarly controlled experiments to increase the number of biological replicates

would, in turn, increase confidence in our ML results. These findings offer valuable information for further studies into the impact of spaceflight on female astronaut health, reiterates well-established roles between spaceflight and breast cancer risk, and provides a straightforward ML approach to leverage a vast array of unexplored data.

Acknowledgments

The authors wish to acknowledge the JMIR reviewers who generously shared their time and expertise to provide invaluable feedback to improve this manuscript. The authors sincerely appreciate the opportunity to have openly discussed this manuscript with them.

Funding

This manuscript is the product of citizen science. No funding was made available for this research.

Data Availability

The notebook for this research is available at [54]. The OSD-511 dataset is available at [55].

Authors' Contributions

JC designed the experiments and wrote most of the manuscript. TZ and JY organized the efforts of the student researchers (AA, AR, AM, AF, KS, SL, WG, AL) who explored alternative approaches to processing the data and validated the references. The ensemble approach was conceived with SC; using linear decision boundary classifiers for ease of interpretation was conceived by MSC. All authors proofread the manuscript and provided their feedback.

Conflicts of Interest

None declared.

References

1. Nguyen HP, Tran PH, Kim KS, Yang SG. The effects of real and simulated microgravity on cellular mitochondrial function. NPJ Microgravity. Nov 8, 2021;7(1):44. [doi: 10.1038/s41526-021-00171-7] [Medline: 34750383]
2. Beheshti A, Miller J, Kidane Y, Berrios D, Gebre SG, Costes SV. NASA GeneLab project: bridging space radiation omics with ground studies. Radiat Res. Jun 2018;189(6):553-559. [doi: 10.1667/RR15062.1] [Medline: 29652620]
3. Beheshti A, Cekanaviciute E, Smith DJ, Costes SV. Global transcriptomic analysis suggests carbon dioxide as an environmental stressor in spaceflight: a systems biology GeneLab case study. Sci Rep. Mar 8, 2018;8(1):4191. [doi: 10.1038/s41598-018-22613-1] [Medline: 29520055]
4. Hughson RL, Helm A, Durante M. Heart in space: effect of the extraterrestrial environment on the cardiovascular system. Nat Rev Cardiol. Mar 2018;15(3):167-180. [doi: 10.1038/nrcardio.2017.157] [Medline: 29053152]
5. Comfort P, McMahon JJ, Jones PA, et al. Effects of spaceflight on musculoskeletal health: a systematic review and meta-analysis, considerations for interplanetary travel. Sports Med. Oct 2021;51(10):2097-2114. [doi: 10.1007/s40279-021-01496-9] [Medline: 34115344]
6. Crucian BE, Choukèr A, Simpson RJ, et al. Immune system dysregulation during spaceflight: potential countermeasures for deep space exploration missions. Front Immunol. 2018;9:1437. [doi: 10.3389/fimmu.2018.01437] [Medline: 30018614]
7. Van Ombergen A, Demertzi A, Tomilovskaya E, et al. The effect of spaceflight and microgravity on the human brain. J Neurol. Oct 2017;264(Suppl 1):18-22. [doi: 10.1007/s00415-017-8427-x] [Medline: 28271409]
8. Beheshti A, Chakravarty K, Fogle H, et al. Multi-omics analysis of multiple missions to space reveal a theme of lipid dysregulation in mouse liver. Sci Rep. Dec 16, 2019;9(1):19195. [doi: 10.1038/s41598-019-55869-2] [Medline: 31844325]
9. Mao X, Stanbouly S, Holley J, Pecaut M, Crapo J. Evidence of spaceflight-induced adverse effects on photoreceptors and retinal function in the mouse eye. Int J Mol Sci. Apr 17, 2023;24(8):7362. [doi: 10.3390/ijms24087362] [Medline: 37108526]
10. Drago-Ferrante R, Di Fiore R, Karouia F, et al. Extraterrestrial gynecology: could spaceflight increase the risk of developing cancer in female astronauts? An updated review. Int J Mol Sci. Jul 5, 2022;23(13):7465. [doi: 10.3390/ijms23137465] [Medline: 35806469]
11. Kumar K, Angdisen J, Ma J, Datta K, Fornace AJ, Suman S. Simulated galactic cosmic radiation exposure-induced mammary tumorigenesis in *Apc*$^{Min/+}$ mice coincides with activation of ERα-ERRα-SPP1 signaling axis. Cancers (Basel). Nov 26, 2024;16(23):3954. [doi: 10.3390/cancers16233954] [Medline: 39682141]
12. Helm JS, Rudel RA. Adverse outcome pathways for ionizing radiation and breast cancer involve direct and indirect DNA damage, oxidative stress, inflammation, genomic instability, and interaction with hormonal regulation of the breast. Arch Toxicol. May 2020;94(5):1511-1549. [doi: 10.1007/s00204-020-02752-z] [Medline: 32399610]

13. Nassef MZ, Kopp S, Melnik D, et al. Short-term microgravity influences cell adhesion in human breast cancer cells. Int J Mol Sci. Nov 15, 2019;20(22):5730. [doi: 10.3390/ijms20225730] [Medline: 31731625]
14. Barcellos-Hoff MH, Ravani SA. Irradiated mammary gland stroma promotes the expression of tumorigenic potential by unirradiated epithelial cells. Cancer Res. Mar 1, 2000;60(5):1254-1260. [Medline: 10728684]
15. Mukhopadhyay R, Costes SV, Bazarov AV, Hines WC, Barcellos-Hoff MH, Yaswen P. Promotion of variant human mammary epithelial cell outgrowth by ionizing radiation: an agent-based model supported by in vitro studies. Breast Cancer Res. 2010;12(1):R11. [doi: 10.1186/bcr2477] [Medline: 20146798]
16. Bartsch C, Bartsch H, Peschke E. Light, melatonin and cancer: current results and future perspectives 1. Biol Rhythm Res. Feb 2009;40(1):17-35. [doi: 10.1080/09291010802066983]
17. Malhan D, Schoenrock B, Yalçin M, Blottner D, Relógio A. Circadian regulation in aging: implications for spaceflight and life on earth. Aging Cell. Sep 2023;22(9):e13935. [doi: 10.1111/acel.13935] [Medline: 37493006]
18. Arnold C, Casaletto J, Heller P. Spaceflight disrupts gene expression of estrogen signaling in rodent mammary tissue. MRAJ. 2024;12(3):3. URL: https://esmed.org/MRA/mra/issue/view/162 [Accessed 2025-12-03] [doi: 10.18103/mra.v12i3.5220]
19. Tang J, Fernandez-Garcia I, Vijayakumar S, et al. Irradiation of juvenile, but not adult, mammary gland increases stem cell self-renewal and estrogen receptor negative tumors. Stem Cells. Mar 2014;32(3):649-661. [doi: 10.1002/stem.1533] [Medline: 24038768]
20. Mertens AC, Liu Q, Neglia JP, et al. Cause-specific late mortality among 5-year survivors of childhood cancer: the Childhood Cancer Survivor Study. J Natl Cancer Inst. Oct 1, 2008;100(19):1368-1379. [doi: 10.1093/jnci/djn310] [Medline: 18812549]
21. Monti N, Masiello MG, Proietti S, et al. Survival pathways are differently affected by microgravity in normal and cancerous breast cells. Int J Mol Sci. Jan 16, 2021;22(2):862. [doi: 10.3390/ijms22020862] [Medline: 33467082]
22. Kannan S, Shailesh H, Mohamed H, Souchelnytskyi N, Souchelnytskyi S. A long-term 10G-hypergravity exposure promotes cell-cell contacts and reduces adhesiveness to a substrate, migration, and invasiveness of MCF-7human breast cancer cells. Exp oncol. May 2023;44(1):23-30. [doi: 10.32471/exp-oncology.2312-8852.vol-44-no-1.17270]
23. Sarkar SR, Pampaloni F. In vitro models of bone marrow remodelling and immune dysfunction in space: present state and future directions. Biomedicines. Mar 2022;10(4):766. [doi: 10.3390/biomedicines10040766]
24. Auslander N, Gussow AB, Koonin EV. Incorporating machine learning into established bioinformatics frameworks. Int J Mol Sci. Mar 12, 2021;22(6):2903. [doi: 10.3390/ijms22062903] [Medline: 33809353]
25. Casaletto JA, Scott RT, Myrick M, et al. Analyzing the relationship between gene expression and phenotype in space-flown mice using a causal inference machine learning ensemble. Sci Rep. Jan 18, 2025;15(1):2363. [doi: 10.1038/s41598-024-81394-y] [Medline: 39824847]
26. Feldner-Busztin D, Firbas Nisantzis P, Edmunds SJ, et al. Dealing with dimensionality: the application of machine learning to multi-omics data. Bioinformatics. Feb 3, 2023;39(2):btad021. [doi: 10.1093/bioinformatics/btad021] [Medline: 36637211]
27. Jovic A, Brkic K, Bogunovic N. A review of feature selection methods with applications. Presented at: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO); May 25-29, 2015:1200-1205; Opatija, Croatia. [doi: 10.1109/MIPRO.2015.7160458]
28. Rincy TN, Gupta R. Ensemble learning techniques and its efficiency in machine learning: a survey. Presented at: 2020 2nd International Conference on Data, Engineering and Applications (IDEA); Feb 28-29, 2020:1-6; Bhopal, India. 2020. [doi: 10.1109/IDEA49133.2020.9170675]
29. United States. Health Research Extension Act of 1985. Public Law 99-158. US Statut Large. 1985;99(Title IV Sections 1-12). [Medline: 11686169]
30. Sanders LM, Lopez DK, Wood AE, et al. Celebrating 30 years of access to NASA Space Life Sciences data. Gigascience. Jan 2, 2024;13:giae066. [doi: 10.1093/gigascience/giae066] [Medline: 39283686]
31. Galazka JM, et al. Transcriptional profiling of mammary glands from mice flown on the RRRM-1 mission. NASA GeneLab. Aug 3, 2022. [doi: 10.26030/WDPR-VA45]
32. DeMaris A, Selman SH. Logistic regression. In: Converting Data into Evidence. Springer; 2013:115-136. [doi: 10.1007/978-1-4614-7792-1_7]
33. Singh J, Banerjee R. A study on single and multi-layer perceptron neural network. Presented at: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC); Mar 27-29, 2019:35-40; Erode, India. Mar 2019.[doi: 10.1109/ICCMC.2019.8819775]
34. Cortes C, Vapnik V. Support-vector networks. Mach Learn. Sep 1995;20(3):273-297. [doi: 10.1007/BF00994018]
35. Dietterich TG. Ensemble methods in machine learning. Presented at: Multiple Classifier Systems, in Lecture Notes in Computer Science; Jun 21-23, 2000:1-15; Cagliari, Italy. [doi: 10.1007/3-540-45014-9_1]

36. ShinyGO 0.85.1. URL: http://bioinformatics.sdstate.edu/go/ [Accessed 2025-12-23]
37. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. Jan 1, 2000;28(1):27-30. [doi: 10.1093/nar/28.1.27] [Medline: 10592173]
38. Pérez-González A, Bévant K, Blanpain C. Cancer cell plasticity during tumor progression, metastasis and response to therapy. Nat Cancer. Aug 2023;4(8):1063-1082. [doi: 10.1038/s43018-023-00595-y] [Medline: 37537300]
39. Kourtis N, Tavernarakis N. Cellular stress response pathways and ageing: intricate molecular relationships. EMBO J. May 17, 2011;30(13):2520-2531. [doi: 10.1038/emboj.2011.162] [Medline: 21587205]
40. Ferrucci L, Fabbri E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. Nat Rev Cardiol. Sep 2018;15(9):505-522. [doi: 10.1038/s41569-018-0064-2] [Medline: 30065258]
41. Azzam EI, Jay-Gerin JP, Pain D. Ionizing radiation-induced metabolic oxidative stress and prolonged cell injury. Cancer Lett. Dec 31, 2012;327(1-2):48-60. [doi: 10.1016/j.canlet.2011.12.012] [Medline: 22182453]
42. Rizzo AM, Corsetto PA, Montorfano G, et al. Effects of long-term space flight on erythrocytes and oxidative stress of rodents. PLoS One. 2012;7(3):e32361. [doi: 10.1371/journal.pone.0032361] [Medline: 22412864]
43. Rakhshandehroo M, Knoch B, Müller M, Kersten S. Peroxisome proliferator-activated receptor alpha target genes. PPAR Res. 2010;2010:1-20. [doi: 10.1155/2010/612089] [Medline: 20936127]
44. Dai DF, Chiao YA, Marcinek DJ, Szeto HH, Rabinovitch PS. Mitochondrial oxidative stress in aging and healthspan. Longev Healthspan. 2014;3(1):6. [doi: 10.1186/2046-2395-3-6] [Medline: 24860647]
45. Qian Z, Chen L, Liu J, Jiang Y, Zhang Y. The emerging role of PPAR-alpha in breast cancer. Biomedicine & Pharmacotherapy. May 2023;161:114420. [doi: 10.1016/j.biopha.2023.114420]
46. Simeone AM, Tari AM. How retinoids regulate breast cancer cell proliferation and apoptosis. Cell Mol Life Sci. Jun 2004;61(12):1475-1484. [doi: 10.1007/s00018-004-4002-6] [Medline: 15197471]
47. Stoll BA. Linkage between retinoid and fatty acid receptors: implications for breast cancer prevention. Eur J Cancer Prev. Aug 2002;11(4):319-325. [doi: 10.1097/00008469-200208000-00002] [Medline: 12195157]
48. Crowe DL, Chandraratna RAS. A retinoid X receptor (RXR)-selective retinoid reveals that RXR-alpha is potentially a therapeutic target in breast cancer cell lines, and that it potentiates antiproliferative and apoptotic responses to peroxisome proliferator-activated receptor ligands. Breast Cancer Res. 2004;6(5):R546-55. [doi: 10.1186/bcr913] [Medline: 15318936]
49. Plutzky J. The PPAR-RXR transcriptional complex in the vasculature: energy in the balance. Circ Res. Apr 15, 2011;108(8):1002-1016. [doi: 10.1161/CIRCRESAHA.110.226860] [Medline: 21493923]
50. Bougarne N, Weyers B, Desmet SJ, et al. Molecular actions of PPARα in lipid metabolism and inflammation. Endocr Rev. Oct 1, 2018;39(5):760-802. [doi: 10.1210/er.2018-00064] [Medline: 30020428]
51. López-Otín C, Pietrocola F, Roiz-Valle D, Galluzzi L, Kroemer G. Meta-hallmarks of aging and cancer. Cell Metab. Jan 3, 2023;35(1):12-35. [doi: 10.1016/j.cmet.2022.11.001] [Medline: 36599298]
52. Zhang TH, Hasib MM, Chiu YC, et al. Transformer for gene expression modeling (T-GEM): an interpretable deep learning model for gene expression-based phenotype predictions. Cancers (Basel). Sep 29, 2022;14(19):4763. [doi: 10.3390/cancers14194763] [Medline: 36230685]
53. Smith AM, Walsh JR, Long J, et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. BMC Bioinformatics. Mar 20, 2020;21(1):119. [doi: 10.1186/s12859-020-3427-8] [Medline: 32197580]
54. Mammary_final_v3.ipynb. Google Colab. URL: https://colab.research.google.com/drive/1ZLB32UQZ_Byc9ja0DpLs30u0VvwuAucu [Accessed 2025-12-23]
55. OSD-511. version 4. transcriptional profiling of mammary glands from mice flown on the RRRM-1 mission. NASA OSDR. URL: https://osdr.nasa.gov/bio/repo/data/studies/OSD-511 [Accessed 2025-12-23]

Abbreviations

AI: artificial intelligence
cnd: condition
FLT: flight
GC: ground control
KEGG: Kyoto Encyclopedia of Genes and Genomes
LR: logistic regression
ML: machine learning
NASA: National Aeronautics and Space Administration
OLD: old
PCA: principal component analysis
PPAR: peroxisome proliferator-activated receptor

RNA-seq: RNA-sequencing
RRRM-1: Rodent Research Reference Mission 1
RSEM: RNA-Seq by Expectation Maximization
SGD: stochastic gradient descent
SLP: single-layer perceptron
STAR: Spliced Transcripts Alignment to a Reference
SVM: support vector machine
VIV: vivarium
YNG: young

Edited by Amy Schwartz; peer-reviewed by Sylvester Sakilay, Mitchell Collier, Arya Rahgozar, Toba Olatoye, Simon Muhindi Savai, Myron Pulier, Randa Salah Gomaa Mahmoud, Clara Amaka Nkpoikanke Akpan, Sayan Mitra, Julie Moonga; submitted 24.Feb.2025; final revised version received 26.Oct.2025; accepted 26.Nov.2025; published 14.Jan.2026

Please cite as:
Casaletto JA, Zhao T, Yeung J, Lee A, Ansari A, Fry A, Mishra A, Raj A, Sun K, Lendahl S, Guan W, Cline MS, Costes SV
Machine Learning Ensemble Investigates Age in the Transcriptomic Response to Spaceflight in Murine Mammary Tissue: Observational Study
JMIRx Bio 2026;4:e73041
URL: https://bio.jmirx.org/2026/1/e73041
doi: 10.2196/73041