Investigating a Microgravity Earth Analog and the Effectiveness of a Nutritional Countermeasure with Machine Learning

Ritu Shyam

A thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the Master of Computer Science and Concentration Applied Artificial Intelligence

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

© Ritu Shyam, Ottawa, Canada, 2025

Abstract

As space travel advances, it becomes increasingly important to develop effective countermeasures against the negative consequences of microgravity exposure experienced by space travelers, such as bone density loss, muscle atrophy, cardiovascular issues, and the reactivation of dormant viruses. This work presents a machine-learning-based investigation of a nutritional countermeasure designed for and evaluated within a bed rest Earth analog. To improve the effectiveness of the data analysis, multiple transformation and normalization techniques are applied to the original dataset. Various classification models are then employed to assess the efficacy of the countermeasure and to identify physiological and molecular changes corresponding to different phases of the bed rest model. *The analysis does not reveal any measurable and consistent effect attributable to the countermeasure*. To demonstrate that the modest sample size of the bed rest data does not impose insurmountable limitations on the detectability of potential countermeasure effects, synthetic datasets of varying sizes are also generated and analyzed similarly.

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor, Dr. Tom Cesari, for his unwavering support, insightful guidance, and invaluable feedback throughout this research.

I am also deeply thankful to Dr. Odette Laneuville for providing the gene expression and biological data, as well as for her expert insights, which significantly enhanced the biological components of this work.

Special thanks to Alexia Walker for her assistance with the Gene Ontology (GO) analysis.

I also appreciate the support and encouragement from my colleagues and friends, whose thoughtful discussions contributed to the development of this study.

Although all biological statements were written with guidance from Dr. Odette Laneuville and Alexia Walker, I acknowledge that I do not have formal training in molecular biology, and therefore some of my biological interpretations may have limitations.

I also acknowledge the use of artificial intelligence (AI) tools during the writing of this thesis. AI was used as a supportive resource for tasks such as formatting references and automating citation placement, paraphrasing and refining sentences for writing grammatically correct sentences with coherent paragraphs, and assisting in debugging code.

Importantly, AI was not used to generate scientific content or results but rather served as a supporting tool to enhance editing, formatting, structuring and revising, throughout the writing process.

Data Source and Collaborations

1. **Core biological data (obtained from the European Space Agency)**

   Health indices, termed international standard measurements (ISM), were collected from participants of the Cocktail bed rest study by the European Space Agency (ESA) who designed, supervised and funded the study at the Institute of Space Medicine and Physiology (MEDES-IMPS). All medical tests were supervised by ESA according to a schedule established by the Agency. Dr. Laneuville was given access to ISM results from individual participants of the Cocktail bed rest study via the competition AO-2023-BedrestCoreData (Announcement of Opportunity), soliciting for experiments to perform retrospective analyses of the anonymized data of seven bed rest studies.

2. **Gene expression data (collected by Dr. Laneuville)**

   Gene-expression data correspond to the transcriptomic response of leukocytes from the participants of the Cocktail bed rest study characterized by the research team of Dr. Laneuville. Study design, blood samples collection, RNA composition previously published in [22]. Preparation of the gene expression data reporting expression values used in the analysis corresponding to individual transcript at 10 time points for the 20 participants was performed by the group of Dr. Laneuville and used in the current analysis.

3. **Gene Ontology (GO) analysis**

   Alexia Walker participated in the GO analysis of the gene features and discussion of the biological interpretations of the findings.

Contents

List of Figures

List of Tables

List of Appendices

Acronyms

Acronym	Full Form
AI	Artificial Intelligence
ALR	Additive Log-Ratio
AUC	Area Under the Curve
BDC	Baseline Data Collection
CNN	Convolutional Neural Network
CLR	Centered Log-Ratio
CV	Cross-Validation
DESeq2	Differential Expression Sequencing (v2)
DL	Deep Learning
ESA	European Space Agency
F1	F1 Score
FN	False Negative
FP	False Positive
HDT	Head-Down Tilt
ILR	Isometric Log-Ratio
KNN	k-Nearest Neighbors
knn n	k-Nearest Neighbors with k = n
LOSO	Leave-One-Subject-Out

Acronym	Full Form
MCC	Matthews Correlation Coefficient
ML	Machine Learning
NN	Neural Network
PCA	Principal Component Analysis
R	Reambulation
RF	Random Forest
rf n	Random Forest with n_estimators = n
SHAP	SHapley Additive exPlanations
SMAD	Suppressor of Mothers Against Decapentaplegic (a family of proteins)
SVM	Support Vector Machine
TP	True Positive
TN	True Negative
TGF-β	Transforming Growth Factor Beta
VAE	Variational Autoencoder

Introduction and Literature Review

Motivation and biological questions

Space exploration presents space travelers with various challenges that significantly impact their health. NASA identified five hazards associated with space travel: prolonged exposure to microgravity, cosmic radiation, confinement and isolation, distance from Earth, and hostile/closed environments [1] [19]. Space hazards collectively lead to a wide array of adverse physiological effects and overall have a negative impact on health limiting the performance of astronauts while in space. Negative effects on physiological systems reported include: muscle atrophy and loss of muscle strength due to mechanical unloading [2] [3], deterioration in bone density and structural integrity [6], cardiovascular deconditioning marked by reduced cardiac function and arterial stiffness [7] [8], hematological alterations like anemia caused by increased red blood cell destruction [9], metabolic disruptions manifesting as altered lipid profiles [10] and adipose tissue dysfunction [11], cognitive and neurobehavioral impairments including decreased reaction times and attention deficits [12], compromised immune function characterized by altered leukocyte counts, inflammation and infections [13], and ocular disturbances involving retinal folds and optic nerve sheath distension [14] [15] [62]. Maintaining astronauts' health is crucial for the success of long missions and requires understanding of the biological response to the space environment.

The investigation of physiological acclimation in the spaceflight environment poses considerable logistical, technical, and analytical obstacles, which inherently constrain the scope, depth, and reproducibility of research. A fundamental limitation stems from the restricted access to space, a consequence of significant financial investment, tightly controlled mission objectives, and rigorous astronaut candidacy requirements, ultimately yielding exceptionally small study populations [16]. This scarcity of space travelers and biological samples combined with differences in the duration of mission constrains statistical power, impedes the generalization of results, preventing the detection of true effects. Furthermore, the unique microgravity environment introduces practical and technical limitations in the execution of complex biological measurements and laboratory analyses. Constraints on payload volume, power supply, equipment weight, and the need to minimize astronaut workload significantly restrict the type and frequency of physiological assessments that can be performed onboard [18]. In addition, the collection of biological data in

orbit is further challenged by limited opportunities for sample storage, preservation, and return to Earth [18]. For example, bone strength, which is reduced after long missions, is detected by dual X-ray absorptiometry scans to assess bone mineral density in clinical settings [6]. The precise identification and characterization of different types of immune cells in blood, as well as their functioning, requires flow cytometry techniques not adapted to the space environment. Moreover, monitoring of multiple health parameters; vital signs, blood analyses, cognitive functions, etc., typically yield high-dimensional data from a small number of subjects, sampled longitudinally, which introduces analytical challenges due to temporal autocorrelation, heterogeneity, and frequent missing data points [19]. These challenges necessitate the development of sophisticated bioinformatic approaches for data preprocessing, integration, and interpretation for assessing astronauts' health and for detecting early signs of dysfunction [19].

Considering these barriers, Earth-based analogs such as head-down tilt (HDT) bed rest models and water immersion serve as indispensable platforms for investigating the physiological response to microgravity in a controlled environment [16] [17]. The participants are selected, and the bed rest models are designed in a standardized way [16]. In the HDT model, participants lie in bed with their heads tilted downwards at –6 degrees for 24 hours a day, 7 days a week. The bed rest position simulates microgravity-induced mechanical unloading and a shift of the blood volume towards the head; two effects encountered in the space environment by astronauts. The blood shift towards the head induces a displacement of fluid from the vascular system to the interstitium (fluid-filled space that surrounding cells, blood vessels, and organs) resulting in a 10%-15% decline in plasma volume [63]. The controlled environment, nutritional control, continuous monitoring, and standardized data collection across multiple systems remedy some of the limitations associated with the study of astronauts. Unlike spaceflight, the bed rest Earth analogue enables the recruitment of larger and more diverse participant cohorts, repeat experiments systematically, and complex design and test intervention protocols aimed at mitigating the negative effects of space that would be impracticable in orbit [19]. Microgravity analogues cannot fully simulate all features of the space environment, cosmic radiation exposure, complete gravitational unloading, distance from Earth but provide an avenue to study the effects of microgravity and isolation on humans [55]. Importantly, bed rest is crucial for intervention testing and risk assessment in space health research [20] [21]. A major benefit of utilizing the HDT bed rest model is the feasibility of conducting intervention studies designed to mitigate or counteract physiological deterioration

induced by microgravity. The three main types of interventions so far tested in bed rest studies are: nutrition, exercise, and artificial gravity [20]. One prominent example a nutritional intervention is the "Cocktail" Toulouse bed rest study conducted in 2017 which tested a nutritional formulation (Appendix I: Cocktail Composition) consisting of polyphenols, omega-3 fatty acids, vitamin E, and selenium, specifically developed to counteract oxidative stress and inflammation pathways induced by immobilization and unloading. The Toulouse Cocktail HDT bed rest study took place at the MEDES clinic in France and provided researchers to rigorously evaluate its efficacy across multiple physiological systems simultaneously [2–15]. Results from the Toulouse bed rest study were analyzed and presented in this thesis.

Head-down tilt (HDT) bed rest studies generally produce two types of data: core data and investigator data. Core data includes standard health measurements collected consistently throughout each bed rest from all participants at specific moments; baseline, HDT bed rest, and reambulation. Core data are collected by the medical team in charge of the study and include daily recordings of vitals, blood analytes, fluid balance, and body weight [2–16]. Investigator-specific data are collected by teams of investigators selected by the European Space Agency to monitor specific physiological systems such as bone, muscle, immune, and cognitive functions requiring specialized measurements gathered to monitor the effects of bed rest and reambulation as well as testing the countermeasure introduced in a study by the European Space Agency. The Toulouse Cocktail study is a recent example of the thorough approach applied to ESA-led bed rest studies, providing fundamental assessments of the participants' health along with investigator-specific data covering the musculoskeletal, cardiovascular, hematological, metabolic, neurological, immune, ocular and other systems as well as an assessment of the effectiveness of the nutritional countermeasure [2–15].

Results from individual investigators have been published in the literature and all concluded the lack of significant effect of the nutritional countermeasure. One exception was the minor improvements were observed in lipid regulation [10] and oxidative stress markers in muscle [2], but these did not extend protective benefits to a wider range of physiological functions. In addition, authors reported the bed rest and reambulation's effects on the physiological system of interest to them. The systemic effects of microgravity-induced by HDT bed rest and countermeasure remain to be evaluated and require an integrated approach for the simultaneous assessment of both core

data and investigators' generated data with the objective to better understand the systemic effects of microgravity and design of effective measures to counteract the negative effects of space on health. A related research challenge is the prediction of resource-intensive physiological metrics using more readily accessible measurements or cross-system data. In particular, immune profiling is characterized by omics data including transcripts and proteins, for the assessment of immune functions and requiring technology and analysis not routinely used in clinics. Conversely, core data like vital signs and blood chemistry are technically easy to monitor and standardize across studies. The Toulouse study began exploring whether transcriptomic and proteomic signals could be correlated with broader physiological changes [22]. Such work is critical to determine whether we can reliably infer investigator-level data from core datasets, reducing reliance on resource-intensive assessments, thus improving operational feasibility in both analogue studies and space missions.

The application of machine learning (ML) techniques enables a novel, data-driven approach to unravel complex patterns within high-dimensional biological data, offering potential for more personalized and precise countermeasures in space health. Machine learning has gained traction in aerospace medicine due to its ability to identify subtle, non-linear patterns across large and heterogeneous datasets. Prior studies have employed ML models to monitor astronaut health, predict health risks during spaceflight, and analyze omics data in simulated microgravity conditions. Unsupervised and supervised learning techniques have been instrumental in identifying biomarkers, classifying physiological states, and predicting outcomes of various spaceflight interventions [50] [51]. Our study leverages ML to investigate the biological and physiological effects of a bed rest study simulating microgravity, enriched with a nutritional intervention.

Machine learning was employed in this study not as a black-box prediction tool, but as a structured, hypothesis-driven approach to assess measurable group separability under various biological and experimental conditions. Unlike prior work that primarily relied on enrichment analysis and dimensionality reduction to characterize group-level trends, this study aimed to directly evaluate whether transcriptomic and physiological responses to prolonged bed rest and nutritional intervention were strong and distinct enough to be captured by classification models. Specifically, machine learning provided a practical framework to test three central hypotheses: whether

individuals receiving the antioxidant-rich cocktail could be distinguished from controls, whether different phases of the bed rest protocol (pre, during, and recovery) showed separable molecular or physiological signatures, whether such separability was consistent across both gene expression and non-gene features, and whether the HDTBR participants had recovered by end of the study. By training simple, interpretable models such as K-Nearest Neighbors [28] [29], Random Forest [24] [30], and shallow Neural Networks [31] [32] on different transformed datasets (e.g., DESeq2-normalized-proportions, $\log_2$ [41] [42], CLR [36–38]), the study quantitatively assessed the detectability of biological changes without relying on unsupervised variance-driven approach. This approach provided a way to infer a biological interpretation of the intervention effects, recovery patterns, and day-specific perturbations in a reproducible, model-agnostic manner.

The dataset used in this research originates from a collaborative project with the European Space Agency (ESA) and the primary investigator (PI) responsible for the gene expression data, Dr. Odette Laneuville. The study involves 20 male subjects who underwent 60 days of head-down tilt bed rest (HDT), designed to mimic the fluid shifts that occur in microgravity [2] [55]. Blood-based gene expression data and physiological measurements were collected longitudinally at baseline and across several post-intervention time points. The inclusion of both transcriptomic and physiological datasets presents a comprehensive platform to study systemic responses to spaceflight analog conditions [22].

Choosing this dataset provided us with several advantages. It combines genetic and physiological data across multiple critical timepoints, offering insight into both the immediate and recovery phases of microgravity analog exposure. Additionally, the presence of a nutritional intervention – administered to a randomized subset of participants – enabled us to compare intervention and control groups in a machine learning framework.

One major challenge in working with this dataset was the inconsistency and preprocessing status of the gene expression data. Instead of raw read counts, the data provided were already processed using DESeq2, a popular normalization method [42]. While DESeq2 is robust for differential expression analysis, it limits flexibility for downstream custom preprocessing and imposes challenges in model interpretability. Further, gene count values transformed with DESeq2 may not reflect inter-subject differences effectively, especially in machine learning pipelines.

The physiological dataset posed different challenges: missing values, feature heterogeneity, and outliers. To ensure model robustness, we applied imputation techniques to fill missing values using the values of the nearest timepoint [48]. Outliers were handled using robust scaling to mitigate the influence of extreme values [78]. Basophils, a notably skewed and low-variance feature, was scaled using Min-Max normalization. Features with a large proportion of outliers were treated and features with minimal biological relevance were excluded after consulting domain experts. 53 Non-gene physiological features across 4 categories were included for the analysis as listed in Appendix IV: **List of Non-Gene Physiological Features**

We retained all biologically meaningful features based on clinical and experimental guidelines. This decision preserved interpretability, which is often lost with automated dimensionality reduction techniques. Given the small sample size and complex longitudinal design, retaining domain-relevant features also allowed the integration of prior biological knowledge.

Machine Learning Models for classification

Three main model families were chosen based on their suitability for our dataset characteristics:

1. K-Nearest Neighbors (KNN) – chosen for its simplicity and interpretability, especially under the assumption of homogeneous responses. This model is sensitive to local structures in the data, making it suitable for subject-level classification tasks [28] [29] [100] [101].
2. Random Forests (RF) – selected for its robustness to overfitting and ability to provide feature importance. Its ensemble nature and non-parametric flexibility made it a reliable baseline across datasets with varying noise levels [24] [30] [104] [105].
3. Neural Networks (NN) – employed for their capacity to capture non-linear interactions within high-dimensional datasets [31] [32].

Hyperparameters such as the number of neighbors in KNN, the number of estimators in Random Forest, and the architecture of the Neural Network were not optimized using grid search or cross-validation. Instead, values were selected based on standard heuristics or default settings commonly used in the literature. For instance, the number of neighbors (k) in KNN was chosen using the square root of the sample size heuristic [29], while Random Forest was tested with 50, 100, 150, 250, and 500 estimators to assess model consistency [105]. Neural Network architectures were

kept shallow with default configurations, as the aim was not to maximize classification accuracy but to test whether group separability could be detected by simple, interpretable models [31].

Hypotheses

The study was framed as a classification task to test three biologically grounded hypotheses:

- Hypothesis H1: *The nutritional cocktail intervention had a uniform impact across all participants, reflected consistently in both gene expression profiles and non-gene physiological features.*
- Hypothesis H2: *Biological profiles at the first baseline day are significantly distinct from those at the out-of-phase timepoints.*
- Hypothesis H3: *By the final recovery day, participants return to physiological and genetic states comparable to their baseline measurements.*

These hypotheses were evaluated through repeated classification experiments using leave-one-subject-out (LOSO) cross-validation. For each hypothesis, models were trained on subsets of the data corresponding to relevant timepoints, intervention groups, and datasets (gene or physiological).

Results

We observed that cocktail intervention had no consistent effects on both physiological and genetic expression in the subjects, therefore we reject the first hypothesis. We see a significant difference in subjects in baseline and in later phases of the study, therefore we do not reject the second hypothesis. The subjects are still classifiable with high evaluation metric scores in baseline vs last day of recovery, so we reject the third hypothesis. The results and their discussion are presented and explained in detail in the Results and Discussions section.

Contributions

This thesis presents an applied machine learning approach to assess the impact of a nutritional countermeasure (the cocktail) using physiological and gene expression data collected during a spaceflight analog study. The contribution lies in applying classification models to a unique, interdisciplinary dataset and interpreting the results through a biological lens.

By combining model performance with feature importance analysis, the study offers insights into whether observed classification patterns align with known biological effects of the intervention. The value of this approach lies in the development of data-driven interpretations of complex biomedical experiments and demonstrates the utility of machine learning in interdisciplinary, real-world research settings.

Data Description

Experiment Design

This experiment was part of a larger project called the Cocktail Bed Rest Study, which was designed to understand how the human body reacts to long periods of inactivity that mimic microgravity. Since studying astronauts in space is difficult and expensive, scientists use head down tilt bed rest on Earth to simulate similar effects. In the Cocktail study, healthy male participants lay in bed with their heads tilted downward at a -6 degree angle for 60 days 24 hours a day. This position causes blood to shift from the legs towards the head, just like in space, and reduces the forces on muscles and bones, making it a useful a model for studying the effects of weightlessness and microgravity. In addition, the Cocktail study was designed to test the potential benefits of a countermeasure, a nutritional supplement, on the negative effects of HDT bed rest [2–16].

The study was conducted at the Institute of Space Medicine and Physiology (MEDES-IMPS) in Toulouse, France. A total of 20 healthy male participants were selected based on strict medical and psychological criteria [2–16]. Participants were randomly assigned to either the cocktail or the control group. The study was conducted in two separate groups of 10 participants each, with an 8-month gap between them. Ethical approval for the study was obtained in France and Canada, and all participants gave informed consent. Throughout the experiment, they were closely monitored by medical professionals to ensure their safety.

The experiment was divided into three distinct phases:

- **Baseline Data Collection (BDC)**: 14 days of pre-bed-rest monitoring in the facility to establish baseline physiological and gene expression profiles.
- **Head-Down Tilt (HDT)**: 60-days of bed rest with continuous monitoring; only half the participants received the cocktail supplement during the HDT phase.
- **Reambulation (R)**: 30-day recovery period, with 14 days in-clinic monitoring with physiotherapy and 16 days at home.

Data Collection

Data collected has two major data categories:

1. **Core biological data** (collected by the European Space Agency):
    - **Blood**: red and white blood cell counts, immune markers
    - **Liquid balance**: water intake, urine output, hydric balance
    - **Vital signs**: heart rate, ECG, blood pressure, temperature
    - **Plasma volume**: plasma and fluid content
    - **Nutrition**: food intake records and nutrient tracking

the list of all the features is shown in Figure 9.

2. **Gene expression data** (collected by Dr. Laneuville):
    - Leukocyte (isolated from blood) transcriptome data collected on 10 specific days of the study as shown in Figure 1.
    - Measured using high throughput RNA sequencing technology (described in detail in the next section)

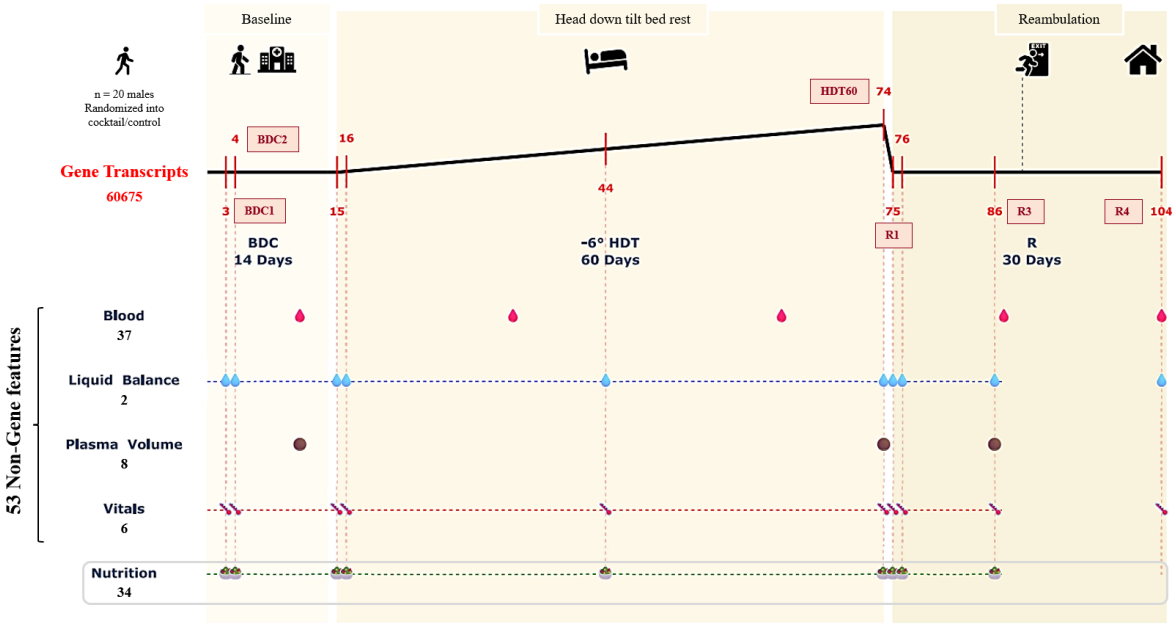


Figure 1: Head down tilt bed rest experiment design.

Gene Expression Data

Gene expression data were derived from blood samples collected in the morning after overnight fasting [22]. The following pipeline was used:

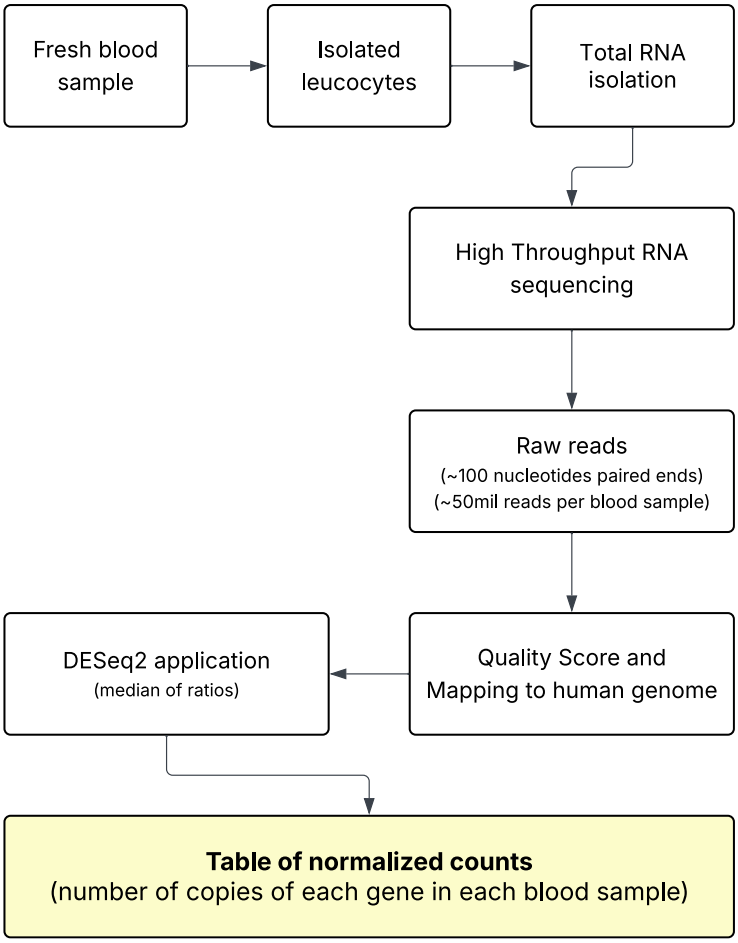


Figure 2: A flowchart showing the process of collecting blood samples to obtaining DESeq2 norm counts.

- **Sample collection**: Blood samples drawn on 10 days across the study (2 times in BDC, 4 HDT, 4 Recovery)
- **White Blood Cells isolation**: White blood cells were isolated using LeukoLOCK™ filters and preserved in RNAlater solution and freeze
- **RNA extraction**: Total RNA was extracted from the isolated white blood cells using the LeukoLOCK™ kit

- **Sequencing**: High-throughput RNA sequencing performed at a depth of ~50 million reads/blood sample. All 200 samples sequenced in the same run (no batch effects) – NovaSeq 6000 platform at Genome Quebec facility in Montreal
- **Quality control, mapping, quantification:** FASTQ files generated from the sequencing facility were analyzed by Dr. Laneuville. In each raw read, nucleotides with high-quality nucleotides (Phred score > 30) kept for further analysis. Mapping and quantification of RNA was done using TopHat and HTSeq tools.
- **DESeq2 pipeline**: Used to normalize the read counts. Genes with low expression (<2 counts) were filtered out.

The output from this pipeline was a **table of normalized gene counts**, which served as the **starting point for the analysis in this manuscript**. These counts are not suitable as direct input to machine learning models due to issues such as high variance, compositionality, and scale sensitivity.

Table 1: Starting point: Table of normalized gene counts

SUBJECTS	**DAYS**	**GENE 1**	**GENE 2**	⋯	**GENE 60675**	**COCKTAIL**
$\boldsymbol{S_i}$	$\boldsymbol{D_{(i,j)}}$	$\boldsymbol{G_g^{(i,j)}}$	$\boldsymbol{G_g^{(i,j)}}$	⋯	$\boldsymbol{G_g^{(i,j)}}$	**0/1**
S_1	$D_{(1,1)}$	$G_1^{(1,1)}$	$G_2^{(1,1)}$	⋯	$G_{60675}^{(1,1)}$	0
⋮	⋮	⋮	⋮	⋯	⋮	⋮
S_1	$D_{(1,10)}$	$G_1^{(1,10)}$	$G_2^{(1,10)}$	⋯	$G_{60675}^{(1,10)}$	0
S_2	$D_{(2,1)}$	$G_1^{(2,1)}$	$G_2^{(2,1)}$	⋯	$G_{60675}^{(2,1)}$	0
⋮	⋮	⋮	⋮	⋯	⋮	⋮
S_2	$D_{(2,10)}$	$G_1^{(2,10)}$	$G_2^{(2,10)}$	⋯	$G_{60675}^{(2,10)}$	0
⋮	⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,1)}$	$G_1^{(20,1)}$	$G_2^{(20,1)}$	⋯	$G_{60675}^{(20,1)}$	1
⋮	⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,10)}$	$G_1^{(20,10)}$	$G_2^{(20,10)}$	⋯	$G_{60675}^{(20,10)}$	1

To address the above limitations, additional **transformations** were applied.

Log2 Transformation

$$\boldsymbol{Log_{2_g}^{(i,j)} = \log_2(G_g^{(i,j)} + 1)} \quad \textbf{(1)}$$

Table 2: Log2 transformation Dataset example structure

SUBJECTS	DAYS	Log_2(GENE 1)	⋯	Log_2(GENE 60675)	COCKTAIL
S_i	$D_{(i,j)}$	$Log_{2_g}^{(i,j)}$	⋯	$Log_{2_g}^{(i,j)}$	0/1
S_1	$D_{(1,1)}$	$\log_2(G_1^{(1,1)} + 1)$	⋯	$\log_2(G_{60675}^{(1,1)} + 1)$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_1	$D_{(1,10)}$	$\log_2(G_1^{(1,10)} + 1)$	⋯	$\log_2(G_{60675}^{(1,10)} + 1)$	0
S_2	$D_{(2,1)}$	$\log_2(G_1^{(2,1)} + 1)$	⋯	$\log_2(G_{60675}^{(2,1)} + 1)$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_2	$D_{(2,10)}$	$\log_2(G_1^{(2,10)} + 1)$	⋯	$\log_2(G_{60675}^{(2,10)} + 1)$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,1)}$	$\log_2(G_1^{(20,1)} + 1)$	⋯	$\log_2(G_{60675}^{(20,1)} + 1)$	1
⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,10)}$	$\log_2(G_1^{(20,10)} + 1)$	⋯	$\log_2(G_{60675}^{(20,10)} + 1)$	1

The log2 transformation is used in gene expression analysis to stabilize variance and address the skewness inherent in count-based data including DESeq2 normalization. Raw or even normalized count values from RNA-seq experiments often span several orders of magnitude, with many low-expression genes and a few high-expression outliers. Applying a log2(x+1) transformation compresses the dynamic range, reducing the influence of extreme values and making the data more suitable for downstream statistical analyses and machine learning algorithms that assume normally distributed input features [41] [42].

One of the key benefits of log2 transformation is variance stabilization, especially when the variance increases with the mean – a typical trait in compositional data. By applying log2(x+1), the variance among highly expressed genes is shrunk, while there are limitations for low-expression gene counts [41]. This helps ensure that no single feature dominates the model simply due to its scale and allows for better detection of differential expression or classification signals. Importantly, the +1 is added to avoid mathematical issues when the count is zero. As a preliminary analysis, log2 transformation was used for its interpretability but it is sensitive to low-count features, where variance can remain high even after transformation [41]. Acknowledging this limitation, we next explored centered log-ratio (CLR) transformation. For future work, the transformations can be extended to incorporate additive log-ratio (ALR) [37] and isometric log-ratio (ILR) transformations [109] for comparative analysis.

CLR Transformation (Centered Log Ratio)

CLR is a transformation applied to compositional data to remove the constant-sum constraint. The formula is:

geometric mean of the gene expression values for sample (i,j)

$$\boldsymbol{GM^{(i,j)} = \left(\prod_{g=1}^{60675} G_g^{(i,j)}\right)^{1/60675}} \quad \textbf{(2)}$$

$$\boldsymbol{CLR_g^{(i,j)} = \log\left(\frac{G_g^{(i,j)}}{GM^{(i,j)}}\right)} \quad \textbf{(3)}$$

Table 3: CLR transformation Dataset example structure

SUBJECTS	DAYS	CLR(GENE 1)	⋯	CLR(GENE 60675)	COCKTAIL
S_i	$D_{(i,j)}$	$CLR_g^{(i,j)}$	⋯	$CLR_g^{(i,j)}$	0/1
S_1	$D_{(1,1)}$	$CLR_1^{(1,1)}$	⋯	$CLR_{60675}^{(1,1)}$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_1	$D_{(1,10)}$	$CLR_1^{(1,10)}$	⋯	$CLR_{60675}^{(1,10)}$	0
S_2	$D_{(2,1)}$	$CLR_1^{(2,1)}$	⋯	$CLR_{60675}^{(2,1)}$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_2	$D_{(2,10)}$	$CLR_1^{(2,10)}$	⋯	$CLR_{60675}^{(2,10)}$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,1)}$	$CLR_1^{(20,1)}$	⋯	$CLR_{60675}^{(20,1)}$	1
⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,10)}$	$CLR_1^{(20,10)}$	⋯	$CLR_{60675}^{(20,10)}$	1

The Centered Log-Ratio (CLR) transformation is particularly suited for compositional data, where the relative proportions of features (in this case, transcripts) matter more than their absolute counts. In RNA-seq data normalized using methods like DESeq2, although counts are adjusted for library size and bias, the data remains compositional in nature – DESeq2 estimates size factors for each

sample based on median ratio method and adjusts counts to correct for sequencing depth. Even after normalization, the counts remain relative. They reflect proportions constrained by the sequencing process, which means if one gene is highly expressed, it reduces the reads assigned to others. CLR transformation mitigates this by converting each gene's expression into a ratio relative to the geometric mean of all gene expressions in the sample and then taking the log of that ratio [36]. This centers the data around zero and ensures the transformed values are scale-invariant, allowing fair comparison across samples with different sequencing depths or library sizes. As such, it addresses a fundamental issue with compositional data: the unit-sum constraint, which artificially introduces correlations and impedes many traditional statistical models. The advantage of CLR transformation lies in its interpretability and its ability to represent the relative expression of genes without being confounded by total read depth. It allows meaningful use of distance-based and geometry-based methods such as K-Nearest Neighbors which are not suitable for raw compositional data.

Proportional Normalization

In addition to commonly used transformations like log2 and CLR, a third approach explored in this study was proportional normalization of DESeq2-normalized counts. In this method, each gene expression value for a given sample was transformed into its proportion of the total gene expression for that sample.

For each sample (i,j), subject S_i on day $D_{(i,j)}$ the proportionally normalized expression of gene g is:

$$\widetilde{G}_g^{(i,j)} = \frac{G_g^{(i,j)}}{\sum_{g=1}^{60675} G_g^{(i,j)}} \quad \textbf{(4)}$$

Table 4: DESeq2 proportional transformation Dataset example structure

SUBJECTS	DAYS	Prop(GENE 1)	⋯	Prop(GENE 60675)	COCKTAIL
S_i	$D_{(i,j)}$	$\widetilde{G}_g^{(i,j)}$	⋯	$\widetilde{G}_g^{(i,j)}$	0/1
S_1	$D_{(1,1)}$	$\tilde{G}_1^{(1,1)}$	⋯	$\tilde{G}_{60675}^{(1,1)}$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_1	$D_{(1,10)}$	$\tilde{G}_1^{(1,1)}$	⋯	$\tilde{G}_{60675}^{(1,10)}$	0
S_2	$D_{(2,1)}$	$\tilde{G}_1^{(2,1)}$	⋯	$\tilde{G}_{60675}^{(2,1)}$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_2	$D_{(2,10)}$	$\tilde{G}_1^{(2,10)}$	⋯	$\tilde{G}_{60675}^{(2,10)}$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,1)}$	$\tilde{G}_1^{(20,1)}$	⋯	$\tilde{G}_{60675}^{(20,1)}$	1
⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,10)}$	$\tilde{G}_1^{(20,10)}$	⋯	$\tilde{G}_{60675}^{(20,10)}$	1

Here, g represents the total number of genes, and the denominator is the sum of all gene expression values for sample (i,j). This transformation results in a gene expression vector for each sample where all values lie in the range [0,1], and the sum across genes equals 1.

This transformation was not applied before log2 or CLR, but rather as an alternative method of normalization to test whether expressing genes as proportions improves model performance. It is especially suited for exploring relative gene importance across samples and helps reducing the influence of absolute magnitude differences across samples. By interpreting each gene $\boldsymbol{G_g^{(i,j)}}$ as a fraction of the sample's total transcriptional activity, this method emphasizes the relative expression pattern of each gene within a sample, aligning conceptually with the assumptions behind compositional data analysis without enforcing the geometric centering of CLR. Since the transformation normalizes each row independently, it also avoids introducing dependencies across samples, making it compatible with models that require i.i.d. (independent and identically distributed) inputs.

The proportional normalization approach is intuitive, interpretable, and computationally simple. Unlike log-based transformations, it avoids mathematical complications associated with zeros. As a result, it can be useful when the focus is on pattern-based classification of samples (e.g., identifying whether a subject belongs to the cocktail group) based on relative gene expression profiles. This transformation was applied as a third normalization strategy in this study, and its performance in binary classification tasks was compared with that of the log2 [42] and CLR [36] transformations to assess its effectiveness.

Non-Gene Biological Features

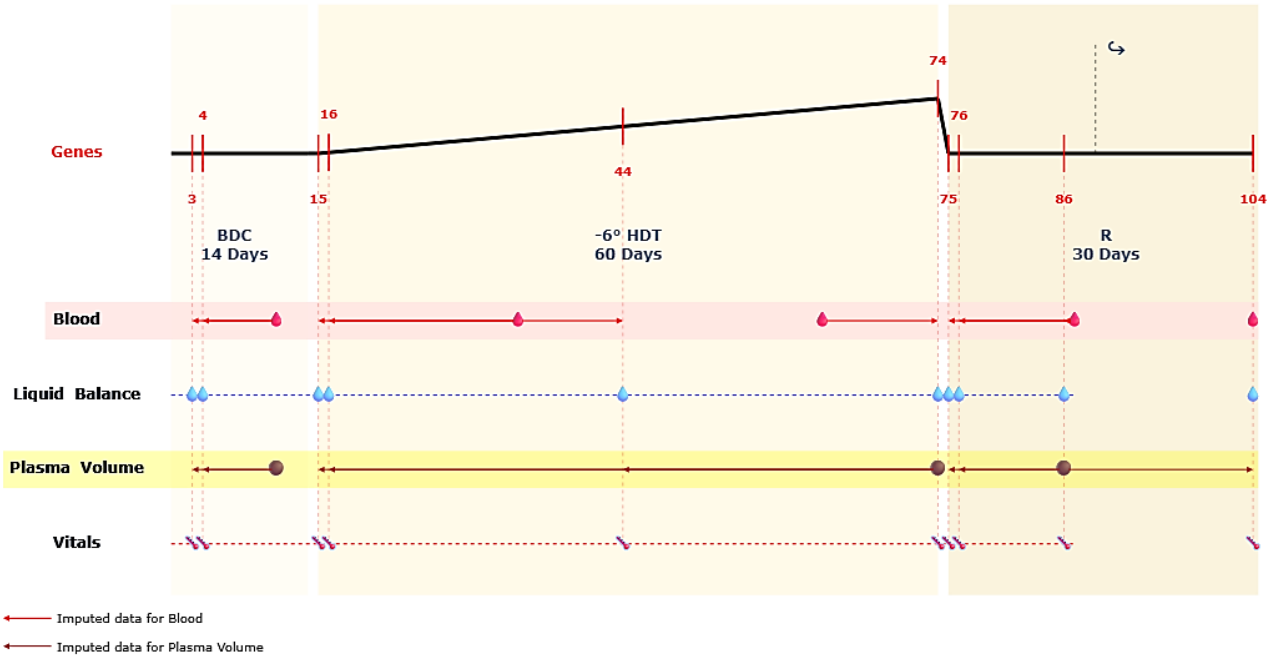


Figure 3: Imputation of missing values in the non-gene pshysiological feature dataset

53 Non-gene physiological features across 4 categories were included for the analysis as listed in Appendix IV: List of Non-Gene Physiological Features. The preprocessing of the non-gene physiological features – missing value imputation, outlier treatment and normalization techniques is discussed in this section.

Missing Value Treatment

Out of 104 days in the experiment, gene profiling was conducted only on 10 specific days [22]. To enable multi-dataset analysis combining both physiological features and gene expression data, the dataset was restricted to these 10 gene-profiling days. However, this introduced missing values for the physiological features on those specific days, since physiological measurements were not always recorded on the same days as gene expression.

To address this, we imputed the missing physiological features using the closest temporal observation from the same study phase. This approach ensures that imputed values are physiologically consistent with the experimental conditions of the corresponding phase: Baseline

Data Collection (BDC), Head-Down Tilt Bed Rest (HDT), or Recovery (R), as illustrated in Figure 3.

This imputation method is inspired by nearest-neighbor imputation [48] [76] in time-series data, adapted to include phase constraints to avoid introducing cross-phase bias, to limit the imputation to the same bed rest phase. It was chosen for the following reasons:

- **Phase-Specific Physiological Relevance**: Physiological changes are expected to follow different trends in each experimental phase; hence, limiting imputation to the same phase improves biological validity.
- **Minimizing Imputation Bias**: By selecting the temporally closest value (rather than strictly using the last observation), we minimize distortion from unrepresentative past or future states [48].
- **Balance Between Accuracy and Interpretability**: The method is simple, traceable, and preserves the integrity of longitudinal signals without overfitting assumptions into missing intervals.

Handling Outliers

Outlier detection and treatment constituted an important part of the preprocessing pipeline. Although physiologically plausible value ranges for each feature were provided by Dr. Laneuville based on domain expertise, we opted for a data-driven approach to ensure consistency and generalizability across all features. Specifically, the Interquartile Range (IQR) method was used to detect extreme values. A data point was considered an outlier if it fell below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$, where Q1 and Q3 represent the first and third quartiles, respectively [43].

Figure 4 shows the number of outliers detected per physiological feature, and the distribution of values – including outliers – is visualized in the boxplots in Figure 5.

To mitigate the influence of outliers during model training, we applied Robust Scaling to all features. This method centers the data around the median and scales it according to the IQR, making it particularly suitable for data with extreme values [78]. One exception was the Basophils feature, which included important zero values that needed to be preserved. For this reason, Min-

Max Scaling was used instead, to normalize the data between 0 and 1 while maintaining the integrity of the zero baseline [43].

This combined approach to outlier handling ensured that the scaling process was both robust to extreme values and sensitive to biological interpretation.

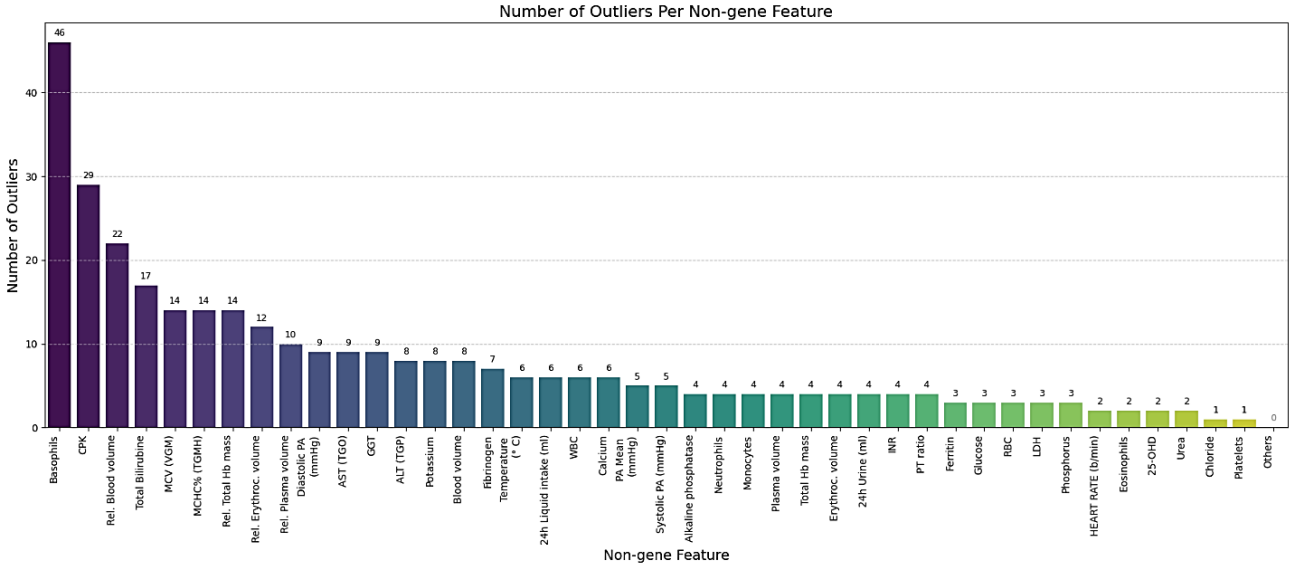


Figure 4: Number of Outliers in the Non-gene Features

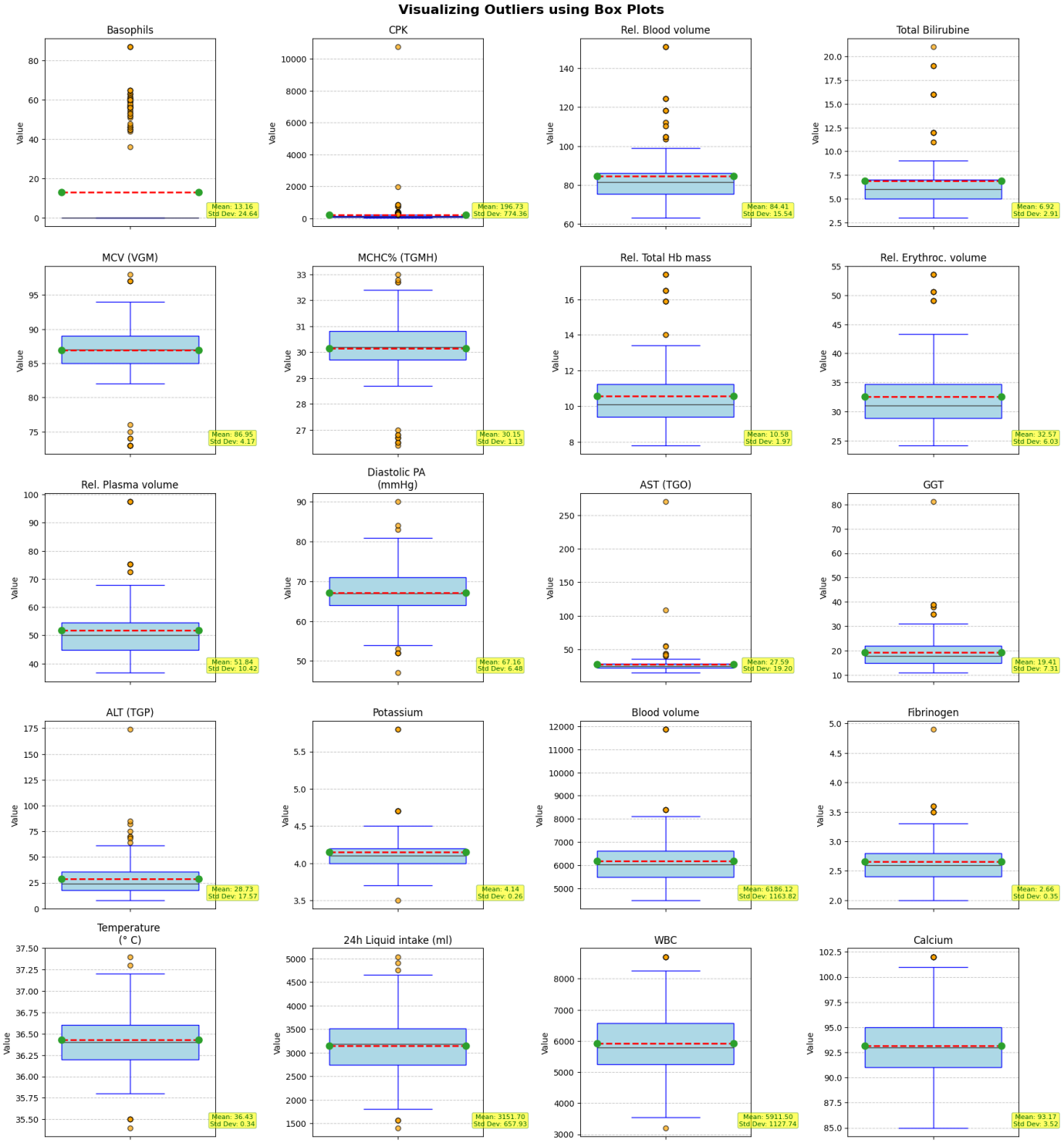


Figure 5: Boxplots of the 20 features with most outliers

Normalization

The normalization of the feature values is explained in the classification section.

Final Non-gene physiological feature dataframe

Table 5: Non-gene Feature Dataset example structure

SUBJECTS	**DAYS**	**Feature 1**	⋯	**Feature 53**	**COCKTAIL**
$\boldsymbol{S_i}$	$\boldsymbol{D_{(i,j)}}$	$\boldsymbol{F_f^{(i,j)}}$	⋯	$\boldsymbol{F_f^{(i,j)}}$	**0/1**
S_1	$D_{(1,1)}$	$F_1^{(1,1)}$	⋯	$F_{53}^{(1,1)}$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_1	$D_{(1,10)}$	$F_1^{(1,10)}$	⋯	$F_{53}^{(1,10)}$	0
S_2	$D_{(2,1)}$	$F_1^{(2,1)}$	⋯	$F_{53}^{(2,1)}$	0
⋮	⋮	⋮	⋯	⋮	0
S_2	$D_{(2,10)}$	$F_1^{(2,10)}$	⋯	$F_{53}^{(2,10)}$	0
⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,1)}$	$F_1^{(20,1)}$	⋯	$F_{53}^{(20,1)}$	1
⋮	⋮	⋮	⋯	⋮	⋮
S_{20}	$D_{(20,10)}$	$F_1^{(20,10)}$	⋯	$F_{53}^{(20,10)}$	1

Synthetic Gene Data Generation via Parametric Perturbation Modeling

To assess how classifier performance scales with biological signal strength in gene expression profiles, we constructed a synthetic dataset using a parametric perturbation-based generative modeling framework. This technique allowed us to simulate biologically realistic control and intervention groups while precisely controlling the degree of class separability. This simulation framework was essential for validating the robustness of our classifiers and understanding the sensitivity of different models to biological effect sizes, without being confounded by small sample sizes or noise inherent in real-world data [47] [84]. In simpler terms, synthetic data lets us create artificial samples that look like real gene expression data. By adjusting how different we make the two classes (control and intervention), we can test if a model is good at detecting even subtle changes. This illustrates that a model could work even in small datasets, as long as there is a sufficiently large difference to detect.

In real-world biological datasets, particularly transcriptomic data, one of the primary challenges is the limited number of samples relative to the number of features (i.e., high-dimensional low-sample-size problem) [47] [50] [51]. This makes it difficult to assess whether poor classification performance is due to inadequate data or truly inseparable classes. Synthetic data generation, where the ground truth is known and controlled, provides a means to isolate and evaluate the learnability of class boundaries under idealized, noise-controlled conditions. It also allows for repeatable and scalable experimentation with different levels of perturbation, helping us draw clearer conclusions about model performance under varying biological signal strengths.

Modeling Framework and Method

We used a single real subject's gene expression vector as the base profile (or "template"). From this, two synthetic classes were generated: a control group, which received only random noise (biological variability), and a cocktail (intervention) group, which received noise plus a systematic shift (biological signal). This simulation approach is grounded in prior literature where gene expression values are modeled using Gaussian distributions to reflect the correlation structure among genes and control the signal-to-noise ratio between classes [86].

Synthetic Data Generation Process:

1. **Input Vector Preparation**:
   A real gene expression vector from one sample is selected and perturbed slightly to avoid zero values, then normalized to sum to 1 to reflect relative gene expression as in the proportional normalization technique of gene data transformation.

2. **Variance Estimation**:
   The variance of each gene is calculated across all samples in the dataset. This variance is used to simulate natural biological variability (noise) in synthetic samples.

3. **Control Group Simulation**:
   For each control sample, noise is drawn from a normal distribution with mean 0 and gene-specific standard deviation and added to the input vector. This simulates natural within-group variation.

4. **Cocktail Group Simulation with Parametric Shift**:
   For cocktail samples, we draw a shift vector from a uniform distribution bounded by a scalar parameter a, representing the *effect size*. The shift vector is used as the mean for the same normal noise generation: This introduces a consistent directional shift, simulating an intervention-related effect.

5. **Post-processing**:
   Negative values (biologically invalid in gene expression) are clipped to zero, and all vectors are re-normalized to preserve the sum-to-one constraint.

6. **Labeling**:
   The resulting vectors are labeled 0 (control) and 1 (cocktail), forming the synthetic dataset for downstream model evaluation.

One real subject's gene profile was used as a starting template for generating synthetic data. For the "control" group, new samples were generated by adding noise to each gene, simulating variation among people. For the "cocktail" group, a shift was also added to simulate the biological change caused by an intervention. To ensure that no values were negative (which could not happen

biologically), the values were clipped. The class labels were finally added so that the complete dataset could be used as an input for the models.

A parameter a governs the magnitude of the shift applied to the cocktail group. Larger values of a simulate stronger biological effects, and vice versa (Figure 6).

This simulation setup allows to systematically test model sensitivity as the biological signal weakens. If models perform well at higher values of the parameter a but fail at lower values of a, it suggests that the classification task is only feasible when the signal is sufficiently strong.

Synthetic Physiological Feature Generation

Following the synthetic simulation of gene expression profiles, a similar perturbation-based generative modeling approach was applied to physiological (non-gene) features in order to assess model sensitivity across different sample sizes.

In contrast to gene expression profiles, which are compositional (i.e., constrained to sum to 1), physiological features such as blood pressure, heart rate, and hydration levels are not bounded by a fixed sum and exhibit absolute, real-valued variation. Therefore, during the generation of synthetic physiological samples, values were clipped at zero to preserve biological plausibility and then robust scaled.

The perturbation strategy involved the addition of zero-mean Gaussian noise to the original feature vector to simulate intra-class (control) variability, and a shift vector scaled by a parameter a to simulate inter-class (cocktail) variation. The parameter a directly modulates the magnitude of synthetic difference between classes, thus allowing precise control over classification difficulty. This method allows us to explore the signal-to-noise landscape of physiological feature spaces under varying perturbation regimes, helping us determine the boundary at classification performance begins to fail.

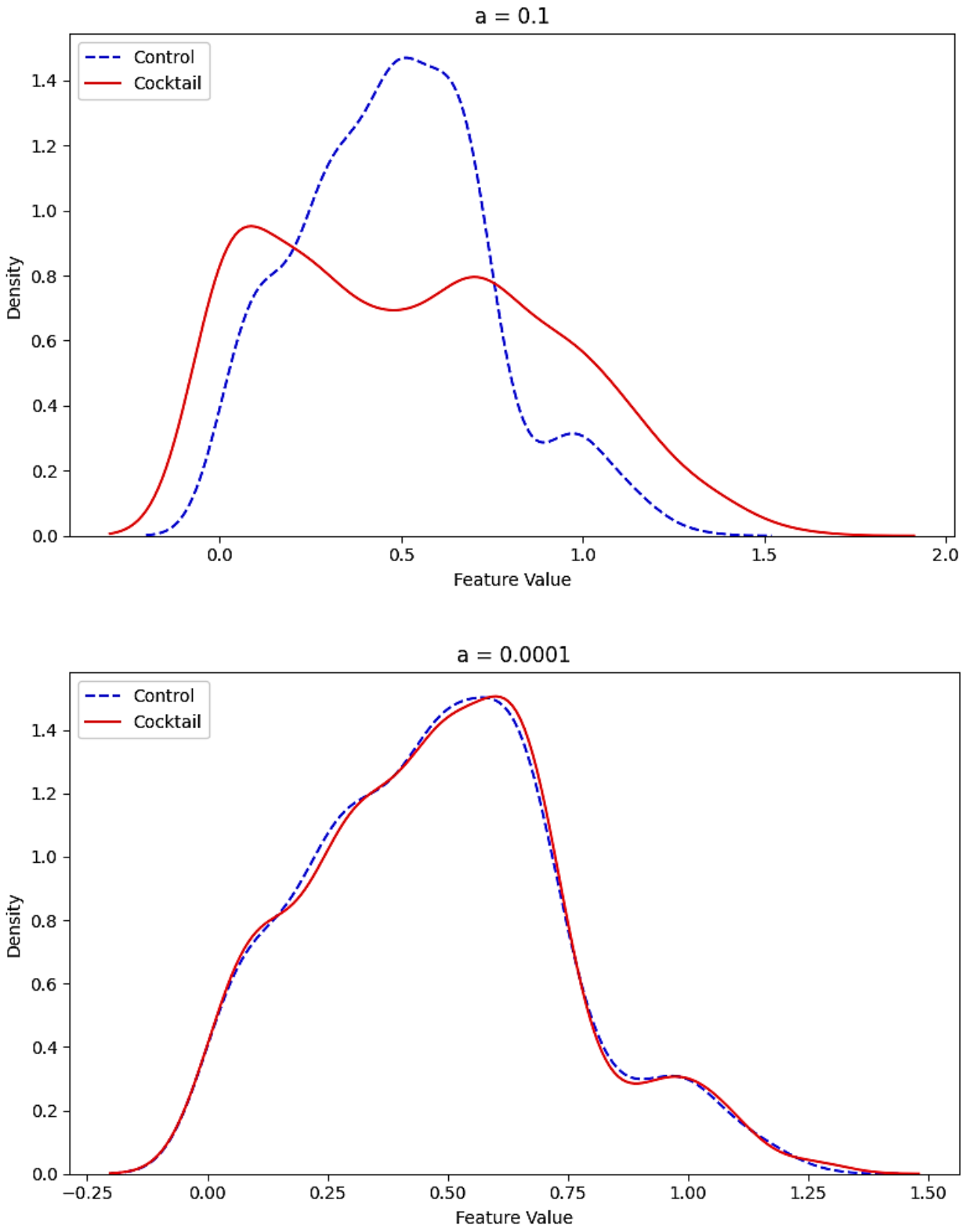


Figure 6: KDE plots for the different perturb "a" value

Table 6: Summary Table of the datasets used as an input for classification into Control vs Cocktail

Dataset	Description	# Sample	# Features	Transformation	Data Type
Log2	Gene expression values log2(x + 1) transformed to stabilize variance.	197	60,675	Log2 transformation	Continuous, normalized
CLR	Centered log-ratio transformed gene expression data to treat compositionality.	197	60,675	CLR transformation	Continuous, normalized
DESeq2 Proportional	DESeq2-normalized gene expression proportions (each row sums to 1).	197	60,675	DESeq2 + proportional normalization	Continuous, compositional
Non-Gene Biological Features	Standard physiological measurements (blood, plasma, vitals, nutrition, etc.).	200	53	Feature-specific (Robust/MinMax)	Continuous, scaled
Synthetic Gene Features	Generated gene expression profiles using perturbation-based modeling	20 / 200 / 2000 / 20000	60,675	DESeq2 + proportional normalization	Continuous, compositional
Synthetic Non-Gene Features	Generated physiological data by introducing realistic variability across features for each group.	20 / 200 / 2000 / 20000	53	Feature-specific (Robust/MinMax)	Continuous, scaled

The datasets used for classification into control and cocktail groups to test the effectiveness of the intervention are described in Table 6. The three gene expression datasets – Log2, CLR, and DESeq2 proportions – contain 197 data points, reflecting only the days when gene sequencing was performed. In contrast, the non-gene dataset includes 200 data points across 53 physiological features, scaled using RobustScaler or MinMaxScaler depending on feature type. All datasets include SUBJECTS and DAYS columns for temporal and subject-level alignment. Missing values were imputed within the same phase. The datasets were carefully aligned such that classification could be performed without introducing temporal mismatch or subject-level inconsistencies.

Table 7: Summary Table of the datasets used as an input for classification into Day-pair classification

SUBJECTS	Dataset	Features	# Sample	# Features	DAY_LABEL
S_i	**Log2**	$Log_2{}_g^{(i,j)}$	40	60,675	0/1
S_i	**CLR**	$CLR_g^{(i,j)}$	40	60,675	0/1
S_i	**DESeq2 Proportional**	$\widetilde{G}_g^{(i,j)}$	40	60,675	0/1
S_i	**Non-Gene Biological Features**	$F_f^{(i,j)}$	40	53	0/1

Each day-pair classification task was designed to compare samples from two specific time points (e.g., BDC1 vs R30), with one sample per subject per day. As a result, each classification dataset consisted of 40 data points – 20 subjects measured at each of the two selected days, forming a balanced binary classification task with 20 samples per class (Table 7). This structure was consistent across all input types: non-gene physiological features, log2-transformed gene expression, CLR-transformed gene expression, and DESeq2-normalized gene proportions. The feature space varied by dataset, with 53 continuous variables in the non-gene dataset and 60,675 gene features in each of the gene datasets. For each data point, the binary target variable, referred to as DAY_LABEL, was assigned a value of 0 if the sample corresponded to the earlier day in the pair and 1 if it belonged to the later day. The same set of subjects and time points was used across all datasets to ensure consistent alignment and comparability. Transformations were applied only to the gene datasets, with log2 transformation used to stabilize variance, CLR to address compositionality, and DESeq2 proportions to normalize total expression levels. Non-gene features were scaled using either robust scaling or min-max normalization depending on the distribution characteristics of each feature. This uniform preprocessing and structure enabled fair comparative classification across different input types while maintaining biological interpretability.

Classification

This study employed supervised machine learning to classify samples based on known labels present in the dataset. The task was framed as classification, where the objective was to assign input data to predefined categories. The datasets used for this study are summarized in the Dataset Summary Tables (Table 6 and Table 7). The analysis was structured around three major classification tasks that explored the biological impact of the nutritional countermeasure and systemic changes across different timepoints, summarized in Table 8.

Table 8: Overview of the classification tasks

Classification Task	Description
Task 1: Cocktail vs Control Classification	**Sample size:** 200 **Objective:** H1: Cocktail leads to consistent physiological/gene differences from Control **Data Splitting:** Leave one subject out (LOSO) cross validation
Task 2: Timepoint-Based (Day-Pair) Classification	**Sample size:** 40 **Objectives:** H2: Baseline is separable from later study phases H3: Subjects return to baseline by R4 **Data Splitting:** 5-fold stratified split
Task 3: Synthetic Data Classification	**Sample size:** 20, 200, 2000, 20000 **Objective:** Understand data limitations **Data Splitting:** 80-20 train-test split

Task 1: Cocktail vs Control Classification

Hypothesis

The nutritional cocktail intervention had a uniform impact across all participants, reflected consistently in both gene expression profiles and international standard physiological (non-gene) features.

Biological Rationale

This binary classification task was designed to evaluate whether the countermeasure elicited consistent physiological or gene expression differences across subjects. A high classification score would suggest a strong and uniform response to the intervention, while near-random performance (accuracy ~ 0.5) would indicate minimal or inconsistent effects.

Evaluation Metric

Accuracy was selected as the principal evaluation metric for this task due to the balanced nature of the classes. The formula for accuracy is:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{5}$$

Where:

- TP = True Positives
  The number of instances that were **correctly predicted as positive** by the model. These are cases where the actual class is positive, and the model also predicted positive.
- TN = True Negatives
  The number of instances that were **correctly predicted as negative** by the model. These are cases where the actual class is negative, and the model also predicted negative.
- FP = False Positives
  The number of instances that were **incorrectly predicted as positive** by the model. These are cases where the actual class is negative, but the model predicted positive. (Also called Type I Error.)

- FN = False Negatives

  The number of instances that were **incorrectly predicted as negative** by the model.

  These are cases where the actual class is positive, but the model predicted negative.

  (Also called Type II Error.)

This choice is further justified by the design of the Leave-One-Subject-Out (LOSO) cross-validation strategy. Since each test fold contains data exclusively from one subject (who belongs to one class), alternative metrics such as precision, recall, F1-score, or Matthews Correlation Coefficient (MCC) become unstable or undefined [25]. For instance, if a subject is misclassified entirely, these metrics reduce to zero or cannot be computed. Therefore, accuracy provides a more stable, interpretable, and suitable metric for this experimental design.

Cross-Validation and Data Splitting

We applied **Leave-One-Subject-Out Cross-Validation (LOSO-CV)**. In each fold, one subject was held out for testing (n = 10), while models were trained on the remaining 19 subjects (n = 190). All preprocessing steps (i.e., normalization) were applied post-split to prevent data leakage. Gene data were standardized using Standard Scaler, while physiological features were scaled using Robust Scalar and Min-Max Scaler.

Scaling

To ensure comparability across features and improve model performance, appropriate scaling techniques were applied to the different datasets in the study.

1. **Gene Expression Data (All Transformations: log2, DESeq2, CLR) — Standard Scaler**

   The gene expression datasets were scaled using the Standard Scaler, which standardizes features by removing the mean and scaling to unit variance:

$$X_{scaled} = \frac{X - \mu}{\sigma} \quad (6)$$

   This approach can be used in gene expression analysis where values (especially post-transformation) are approximately normally distributed or expected to span a wide range around zero. Standardization ensures that features with higher variance do not dominate model training and that all gene measurements contribute equally. As these datasets were already transformed (e.g., via log2 or CLR), outlier influence was less prominent, justifying the use of standard scaling.

2. **Physiological Features (Non-Gene Data) — Robust Scaler**

   All physiological, non-gene features were scaled using the Robust Scaler, which centers data around the median and scales according to the interquartile range (IQR):

$$X_{scaled} = \frac{X - Median(X)}{IQR(X)} \quad (7)$$

   This method was selected due to its robustness to outliers, which are common in biological signals and clinical measurements. Unlike the Standard Scaler, which is highly sensitive to extreme values, RobustScaler minimizes the influence of outliers and preserves the structure of the bulk of the data. This scaling method ensures that features are comparable while reducing the risk of distortion from a few anomalous measurements, which is especially important for real-world physiological data.

3. **Exception – Basophils: Min-Max Scaler**

   For Basophils, Min-Max Scaling was used to transform the values into a fixed range between 0 and 1:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (8)$$

   This choice was made to preserve the 0 values present in the measurements, which may hold biological significance or indicate non-detection. Unlike other scalers, Min-Max

Scaling retains the relative spacing and distribution within a bounded interval, making it particularly useful when zero values must be preserved and distinguished from low, non-zero values.

Task 2: Timepoint-Based (Day-Pair) Classification

Hypothesis

We hypothesized that:

- Within-phase comparisons (e.g., BDC1 vs BDC2) would yield low classification accuracy due to minimal biological change.
- Cross-phase comparisons (e.g., BDC1 vs HDT60) would yield higher classification performance scores due to significant physiological or genetic alterations.
- If BDC1 and R4 are indistinguishable, this would indicate recovery.

Biological Rationale

Timepoint classification allows us to assess how biological features change over time and whether the biological state of subjects evolve predictably through intervention and recovery.

Day-Pairs Compared

- **BDC1 vs BDC2:** Compares two baseline measurements taken before the intervention to assess natural day-to-day variability.
- **BDC1 vs HDT60:** Compares pre-intervention baseline to the final day of 60-day bed rest to evaluate the cumulative effect of prolonged bed rest induced unloading.
- **BDC1 vs R1:** Compares baseline to the first recovery day to capture immediate body responses to reloading.
- **BDC1 vs R3:** Assesses how far recovery has progressed two weeks after reambulation relative to baseline.
- **BDC1 vs R4:** Evaluates whether biological parameters have returned to or diverged from baseline one month after bed rest.

Evaluation Metrics

1. **Accuracy:** Measures the proportion of total correct predictions out of all predictions.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad \text{as shown in Eq. (5)}$$

2. **Precision:** Measures the proportion of correctly predicted positive instances among all predicted positives.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{9}$$

3. **Recall (Sensitivity or True Positive Rate):** Measures the proportion of correctly predicted positive instances among all actual positives.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{10}$$

4. **F1 Score:** Harmonic mean of Precision and Recall, providing a balance between the two.

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})} \tag{11}$$

5. **Matthews Correlation Coefficient (MCC):** Measures the quality of binary classifications, even for imbalanced datasets.

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{12}$$

Cross-Validation and Data Splitting

We employed a 5-fold stratified split (no class imbalance) for training and testing due to smaller available sample sizes. Preprocessing steps were performed only on the training set and then applied to the test set to avoid leakage.

Scaling

Gene data were standardized using Standard Scaler, while physiological features were scaled using Robust Scalar and Min-Max Scaler, similar to Task 1.

Task 3: Synthetic Data Classification

Hypothesis

If the classes (control vs cocktail) have distinct underlying distributions, even small sample size data with small perturbations in data distributions should allow successful classification. However, with lower perturbation values (e.g., a approaching 10^{-60} for gene data and a approaching 10^{-8} for synthetic feature data), classification accuracy should decline.

Biological Rationale

This task simulates data augmentation and explores whether underlying patterns are robust enough to be learned by classifiers even when realistic noise is introduced. It also allows investigation into the minimum sample size required for effective learning.

Evaluation Metrics

We used **Accuracy** as the main evaluation metric, alongside visual inspection of trend curves for multiple sample sizes (20, 200, 2000, 20000). The remaining 4 metrics were also calculated and is presented in Appendix III: Synthetic Data Results. Synthetic experiments help validate our real-world results.

Cross-Validation and Data Splitting

A fixed 80-20 split was used for synthetic datasets, and the process was repeated for each level of perturbation for various sample size of the data.

Scaling

Gene data were standardized using Standard Scaler, while physiological features were scaled using Robust Scalar and Min-Max Scaler, like Task 1.

Classifier Descriptions and Hyperparameters

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, distance-based classification algorithm that assigns a class to a new data point based on the average among its *k* nearest neighbors in the feature space [28] [29]. Because the algorithm does not make any assumptions about the underlying data distribution, it is particularly suitable for biological and physiological datasets where feature relationships may be complex or unknown [100] [101]. KNN operates on the principle that samples close in feature space tend to share the same label.

In this study, KNN was selected to evaluate whether local similarities in physiological and gene expression features were sufficient to distinguish between experimental groups – such as control versus cocktail subjects – or between different temporal phases of the study. The choice of *k*, the number of neighbors, was informed by a common heuristic that recommends selecting *k* as the square root of the total number of samples [29]. Given that the full dataset consisted of 197 samples, and roughly 10 were held out in each fold during Leave-One-Subject-Out (LOSO) cross-validation, the expected optimal *k* was around 14. However, to examine the sensitivity of the model to neighborhood size and better capture local structure, *k* values of 1, 5, and 9 were tested. For timepoint-based classification tasks – where each day pair contained only 40 samples (20 per group) – lower *k* values were deemed more appropriate to avoid excessive smoothing of local distinctions.

Euclidean distance was employed as the metric for determining similarity between samples, as it is widely used for continuous numerical data and is compatible with the normalized features used in this analysis [29]. All features were scaled prior to modeling to ensure fair contribution to distance calculations. Uniform weighting was applied, meaning all neighbors contributed equally to the classification decision, which helped maintain model simplicity and interpretability. Overall, KNN provided a biologically intuitive baseline for assessing the separability of groups within the dataset under minimal modeling assumptions.

Random Forest (RF)

Random Forest is a widely used ensemble learning algorithm based on decision trees [24] [30]. It operates by constructing multiple decision trees during training and outputs the class that is the mode (majority vote) of the predictions of individual trees. The core idea behind this approach is to combine the predictions of several weak learners to produce a more accurate and robust model [45]. This ensemble strategy reduces overfitting typically associated with single decision trees and lowers variance while maintaining interpretability. The trees in the forest are built on different random subsets of the training data and features, which introduces diversity among the learners and helps the model generalize better.

In this study, Random Forest was selected for its good empirical performance on several benchmark datasets [44]. Additionally, Random Forest models are robust to noise [24], which is particularly important in biomedical datasets where biological variability and measurement noise are common. Another compelling advantage is the built-in feature importance metric, which allows identification of the most predictive physiological or genetic variables in a transparent and interpretable manner. This aligns well with the goals of this study, which include not only accurate classification but also gaining biological insights.

A key hyperparameter of the random forest model is n_estimators, which refers to the number of trees in the forest. For the cocktail vs. control group classification, models were evaluated using 50, 100, and 150 trees to explore the impact of ensemble size on stability and accuracy. For day-pair classifications (e.g., BDC1 vs. R1), higher values were tested – specifically 100, 250, and 500 trees – to ensure that sufficient ensemble complexity was available to capture potentially subtle temporal shifts. The choice of values was informed by literature suggesting that model performance tends to plateau after a certain number of trees, with diminishing returns beyond a certain number of trees in many cases [105] [106].

The max_features parameter was left at its default value, which in most implementations (e.g., scikit-learn [57]) selects the square root of the total number of features for classification tasks. This ensures that each tree is trained on a diverse subset of features, enhancing model diversity and generalization. The random_state parameter was fixed at 42 to ensure reproducibility across runs, which is particularly important when comparing model performance across cross-validation folds. Lastly, the bootstrap parameter was set to True to allow sampling with replacement.

Bootstrapping is a critical component of Random Forests, as it allows each tree to see a slightly different version of the training data, further contributing to the ensemble's variance reduction capabilities. Together, these settings ensured that the Random Forest models used in this study were both theoretically sound and empirically tuned to provide stable, generalizable, and interpretable classification results.

Neural Network (NN)

In this study, a shallow feedforward neural network (also known as a multilayer perceptron, or MLP) was implemented as part of the classification framework to explore non-linear patterns within the physiological and transcriptomic datasets. A feedforward neural network is a type of artificial neural network where the connections between the nodes do not form cycles. Data flows in one direction – from input to output – through one or more hidden layers [31]. This architecture is well-suited for modeling complex relationships between features, especially in biomedical data where interactions may not follow linear patterns [93].

The neural network was chosen for its ability to capture intricate, multi-layered interactions that might be missed by simpler models like K-Nearest Neighbors or Random Forests. Unlike decision trees or distance-based methods, neural networks learn internal representations through optimization, making them effective for identifying feature combinations and interactions that are not immediately obvious [31]. This makes them especially useful in high-dimensional spaces, such as those involving gene expression profiles or physiological states influenced by multiple biological systems simultaneously [93].

Another reason for choosing this architecture is its flexibility. Neural networks are inherently adaptable to different feature scales and dimensionalities, whether the input space involves a few physiological measurements or thousands of genes. This adaptability is essential in a study like ours, where datasets vary in size and complexity across different tasks (e.g., intervention classification versus timepoint differentiation).

While deep neural networks with multiple layers and millions of parameters are typically used in large-scale applications, a shallow architecture was deliberately selected in this study to strike a balance between model complexity and the risk of overfitting, especially given the modest sample sizes [32]. The goal was not only to explore classification performance but also to examine whether

the added representational capacity of neural networks could offer any advantage in detecting biologically meaningful distinctions without overly complicating the model.

Overall, the inclusion of a neural network model in this study provided an opportunity to test the robustness of classification results across a spectrum of model complexities, from simple, interpretable models like KNN to more flexible, non-linear learners like neural networks. This ensured that observed patterns were not model-dependent and allowed for a more comprehensive evaluation of separability within the data**.**

Architecture and Hyperparameters

- **Input Layer:** Number of neurons equals number of features. (60675 or 53)
- **Hidden Layers:** Two layers with 128 and 64 neurons.
- **Activation Function:** ReLU for hidden layers.
- **Dropout:** 0.3 between layers to reduce overfitting.
- **Output Layer:** 1 neuron with sigmoid activation for binary classification.
- **Loss Function:** Binary Cross-Entropy.
- **Optimizer:** Adam (learning rate = 0.001).
- **Batch Size:** 8, 20 epochs

Hardware and software specs

All machine learning experiments, data preprocessing, and analysis tasks were conducted on a personal computing system with the following hardware specifications:

- Processor: 13th Gen Intel® Core™ i9-13900HX CPU @ 2.20 GHz
- RAM: 32.0 GB
- System Type: 64-bit operating system, x64-based processor
- Operating System: Windows (university-managed environment)

The software environment used for the implementation included:

- Editor: Visual Studio Code
- Programming Language: Python 3.11.9
- Key Libraries and Frameworks:
    - scikit-learn: for classical machine learning models, evaluation metrics, and model selection utilities
        - RandomForestClassifier, KNeighborsClassifier
        - accuracy_score, precision_score, recall_score, f1_score, matthews_corrcoef
        - train_test_split
    - tensorflow.keras: for implementing the feedforward neural network
        - Sequential, Dense, Dropout, Adam
    - numpy and pandas: for numerical computations and data manipulation
    - matplotlib.pyplot and seaborn: for data visualization

This setup was sufficient for all experimental runs, including data processing, training, validation, and result visualization. No high-performance computing or GPU acceleration was used.

Results and Discussions

Results - Task 1: Classification into control and cocktail

In the classification task distinguishing control from cocktail subjects, accuracy is adopted as the principal evaluation metric. This choice is motivated by the class-balanced nature of the dataset; both classes are nearly equally represented, making accuracy a reliable and interpretable measure of overall model performance [27]. Additionally, the experimental design ensures that each subject retains a consistent class label across all timepoints. In the leave-one-subject-out (LOSO) cross-validation scheme, the model is tested on entirely unseen individuals, closely mimicking real-world generalization. While alternative metrics such as precision, recall, F1-score, and Matthews Correlation Coefficient (MCC) are often used in class-imbalanced scenarios or when type I and type II errors carry unequal importance, they may provide little additional insight here.

More importantly, these metrics may mislead interpretation in the LOSO context. Since each test fold consists entirely of data from a single subject, all predicted labels belong to the same true class. This causes the denominator in precision or recall calculations to become undefined or artificially skewed, particularly if the model predicts the wrong class for the entire subject, resulting in zero true positives. As a result, precision and recall may collapse to zero or be undefined, and F1-score becomes zero despite possibly high classification correctness across folds. Similarly, MCC, while useful for measuring correlation between predictions and true labels, becomes unstable in such small sample test sets where only one class is present per fold [25]. Therefore, using accuracy – which reflects the proportion of correct predictions across all subjects – provides a more stable, interpretable, and sufficient metric for this specific classification task. Reported accuracies in Table 9 represent the macro average across subjects, where performance is computed separately for each subject (fold) and then averaged.

Table 9: Results of Control vs Cocktail Classification across Non-gene and Gene Transformation Datasets for different models

	Non-Gene Features		Transformed Gene Features					
			Log2		CLR		DESeq2	
model_param	mean	std	mean	std	mean	std	mean	std
knn(k=1)	0.52	0.32	0.44	0.22	0.39	0.20	0.55	0.17
knn(k=5)	0.56	0.34	0.46	0.23	0.37	0.15	0.55	0.18
knn(k=9)	0.53	0.36	0.43	0.19	0.33	0.18	0.45	0.14
nn	0.42	0.40	0.41	0.29	0.34	0.27	0.03	0.07
rf(n=50)	0.41	0.40	0.29	0.26	0.31	0.22	0.29	0.31
rf(n=100)	0.36	0.40	0.25	0.29	0.34	0.27	0.22	0.25
rf(n=150)	0.35	0.38	0.24	0.27	0.27	0.25	0.20	0.24

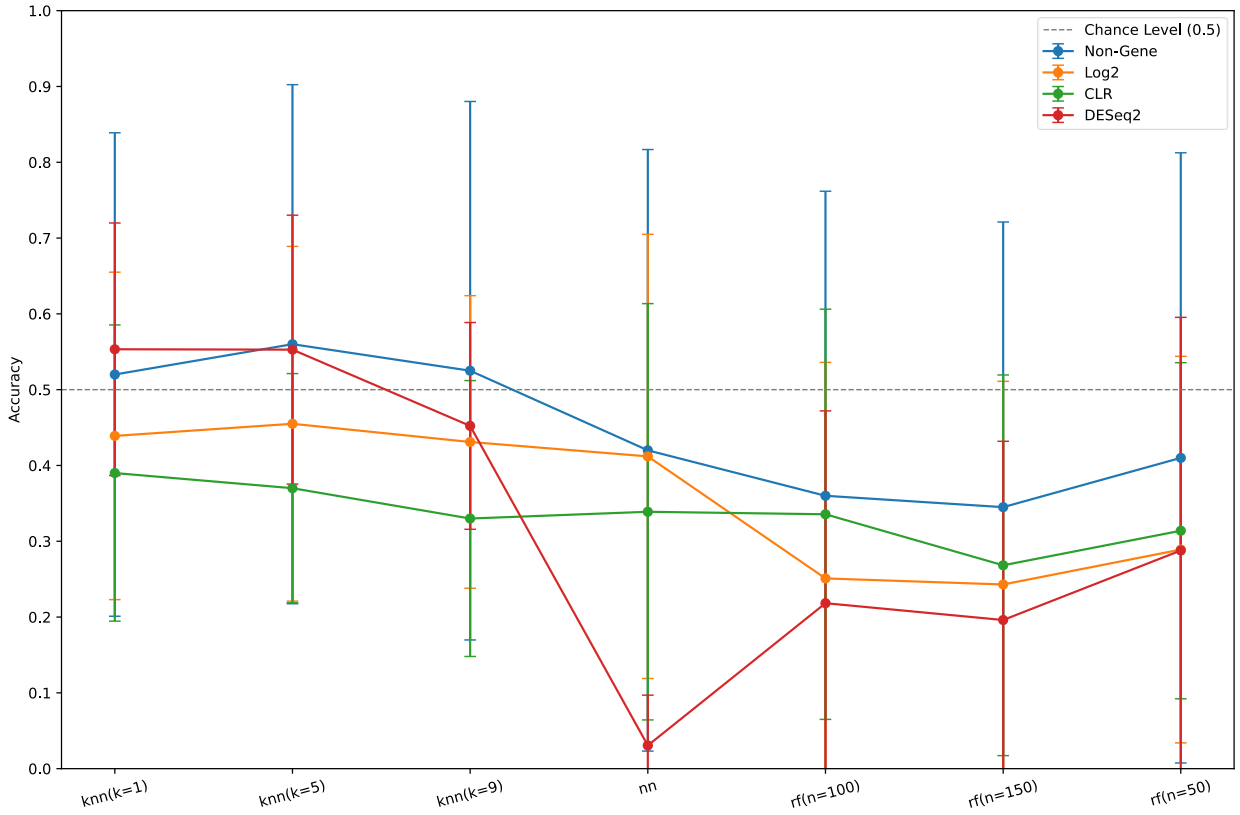


Figure 7: Line graph showing Mean Accuracy and Standard Deviation of Control vs Cocktail Classification

Rejection of Hypothesis H1: No Observable Impact of the Cocktail Intervention

Hypothesis H1: *The nutritional cocktail intervention had a uniform impact on the subjects in both gene expression profiles and international standard physiological (non-gene) features.*

Result: Rejected.

We reject Hypothesis H1. Across all models and input datasets, our classification experiments consistently failed to uniformly distinguish cocktail from control subjects with accuracy significantly greater than that of a random guess. Despite employing multiple machine-learning algorithms, including K-Nearest Neighbors (KNN), Random Forest (RF), and Feedforward Neural Networks (NN), alongside different cross-validation folds and three biologically motivated gene transformation techniques (log2, CLR, DESeq2 proportions), classification performance hovered around 50% moreover with substantial variability (large standard deviation) – see Figure 7. These findings do not support the hypothesis that the nutritional countermeasure had a measurable and consistent effect across the cohort. Therefore, we reject the hypothesis.

One potential concern is whether the low classification accuracy observed might be attributed to limitations or inadequacies in the machine learning models employed. To address this, we implemented a diverse set of algorithms: KNN, a non-parametric method known for its effectiveness in locally clustered datasets; Random Forest, an ensemble learning technique that is robust to high-dimensional data and noise; and a feedforward neural network, which is capable of capturing complex, nonlinear feature relationships. These models span a spectrum of learning paradigms and were tested using appropriate cross-validation techniques. Moreover, the same algorithms performed well in the day-pair classification experiments which will be discussed later. The consistent failure across these varied models to achieve performance beyond random chance suggests that the absence of predictive signal is not due to model limitations.

Another consideration is whether the data transformations applied to the gene expression values might have inadvertently removed meaningful biological signal. To prevent this, we employed three normalization techniques for the gene datasets: log2(x + 1) transformation to stabilize variance, centered log-ratio (CLR) transformation to account for the compositional structure of the data, and proportional normalization of DESeq2 norm counts to transform the absolute values into distributions. Unlike log2 or CLR, proportional normalization does not treat variance or zero

values, which can make certain models like neural networks unstable and lead to poor performance. Nevertheless, each transformation is designed to correct specific statistical artifacts and has been validated extensively in literature [36] [42]. The similarity in classification results across all transformation types provides evidence that the choice of normalization did not obscure any underlying intervention signal, while highlighting that model sensitivity to normalization should be carefully considered.

A key observation across classification experiments was the high standard deviation of accuracy across subjects, which warrants deeper interpretation. This variability is not a confounding factor, but rather a central result. Using Leave-One-Subject-Out (LOSO) cross-validation, each subject is excluded from training and used for testing. This strategy directly measures the model's ability to generalize across individuals. The observed high variance suggests that even if some subjects responded to the intervention, these effects were not generalizable. Thus, the results do not show the intervention did result in a consistent biological and transcriptional profile across participants. This inter-subject heterogeneity is expected in biological systems and further diminishes the likelihood of the cocktail producing uniform effects detectable by machine learning models.

It is also important to assess whether the limited sample size (n = 20) might have prevented reliable detection of intervention effects. While larger sample sizes increase statistical power, they are not the sole determinant of model performance. In fact, our synthetic data experiments which will be discussed later, suggested that even small sample sizes yield high accuracy when class distributions are separable. The failure of classification models to distinguish cocktail from control groups in real data despite identical modeling strategies indicates that the class boundary is not discernible, rather than the sample size being insufficient.

Another consideration is the presence of noise inherent in biological measurements, which could obscure real but subtle effects. Although noise is a recognized factor in omics data, the preprocessing applied for the datasets included robust scaling, domain-informed outlier treatment, and within-phase imputation strategies. These steps are consistent with best practices in biomedical data science and were applied uniformly across all features, as discussed in the introduction. If the cocktail intervention had produced a biologically meaningful effect, we argue that such preprocessing should have preserved sufficient signal for classification, especially given the richness of the data (poor quality blood samples – 3 of the 200 RNA samples – were discarded).

Finally, it is conceivable that the cocktail induced subtle molecular or physiological changes that were not detectable through classification. However, such an interpretation does not align with Hypothesis H1, which explicitly stated that the intervention would produce measurable impacts on gene expression and non-gene features. We argue that if the observed effects are so minimal that they do not alter the distributions of over 60,000 genes or 50+ physiological variables across 10 time points, their practical and biological significance is questionable.

In conclusion, we do not believe Hypothesis H1 is rejected due to methodological shortcomings or data quality issues, but because no consistent, measurable signal of intervention effect was detected. The observed subject-level variability reinforces the need for personalized approaches in future intervention studies.

Results - Task 2: Classification into day-pairs in different phases

To evaluate the impact of genetic and physiological changes across different experimental phases, we performed binary classification between specific day pairs using both non-gene features and gene expression datasets (with three different transformations). While classification between baseline days (BDC1 vs BDC2) consistently failed across all datasets yielding poor performance metrics, the models achieved notably high performance when distinguishing BDC1 from HDT60, as well as from recovery days (R1, R3, and R4). The following section presents a detailed overview of these high-performing day-pair comparisons, with performance metrics summarized inTable 10.

Using the Non-Gene Feature Data

To assess whether physiological changes across different phases of the bed rest study were distinguishable using non-gene features, we conducted binary classification between BDC1 and other key timepoints: HDT60, R1, R3, and R4. The aim was to evaluate two key hypotheses:

- **Hypothesis H2**: *BDC1 is significantly different from the out-of-phase timepoints (HDT60, R1, and R3).*

  Result: Not Rejected.

  The classification models successfully distinguished BDC1 from HDT60, R1, and R3 with high accuracy and other performance metrics (Table 10),indicating that the physiological profiles at these phases were substantially different from baseline. This

supports the hypothesis that bed rest induces measurable deviations from baseline in key physiological features which is still prevalent in the subsequent recovery phases.

- **Hypothesis H3**: *Participants are physiologically recovered to baseline levels by R4.*
  Result: Rejected.
  The classification performance between BDC1 and R4 remained high, suggesting that participants had not fully returned to their baseline physiological state by the end of the recovery period. Despite the passage of time, several features continued to differ from baseline, implying lingering effects of prolonged inactivity or incomplete physiological recovery.

Table 10 below summarizes the classification performance metrics – Accuracy, Precision, Recall, F1 – score, and Matthews Correlation Coefficient (MCC) across all models and datasets for the high-performing day-pair comparisons (BDC1 vs HDT60, R1, R3, R4). Each metric is reported as the mean and standard deviation over 5-fold stratified cross-validation – the metrics were computed on each held-out fold and then macro-averaged across folds (mean ± SD). A fixed random seed was used to ensure reproducibility. In this experimental setup, Stratified K-Fold cross-validation (n_splits = 5) was used to maintain class balance in each fold, ensuring fair evaluation even when data was limited. From the table, it is evident that the classification models achieved high and stable performance when distinguishing BDC1 from HDT60 and recovery timepoints (R1, R3, R4), for the non-gene features dataset. This reinforces the earlier interpretation that these phases exhibit biologically meaningful shifts in physiological characteristics relative to the baseline. The consistently low standard deviation across metrics indicates high stability and generalizability of the classification models for the selected day pairs.

BDC 1 vs BDC 2		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.33	0.14	0.33	0.19	0.45	0.27	0.38	0.22	-0.37	0.30
rf	250	0.38	0.15	0.36	0.21	0.45	0.27	0.40	0.23	-0.26	0.32
rf	500	0.38	0.15	0.36	0.21	0.45	0.27	0.40	0.23	-0.26	0.32
knn	1	0.33	0.14	0.35	0.11	0.40	0.14	0.37	0.11	-0.37	0.30
knn	5	0.38	0.09	0.41	0.06	0.60	0.14	0.49	0.08	-0.29	0.21
knn	9	0.40	0.14	0.40	0.14	0.45	0.21	0.42	0.16	-0.20	0.27
nn		0.33	0.17	0.30	0.21	0.30	0.21	0.30	0.21	-0.35	0.34
BDC 1 vs HDT 60		**Accuracy**		**Precision**		**Recall**		**F1 Score**		**MCC**	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.88	0.09	0.85	0.14	0.95	0.11	0.89	0.07	0.78	0.15
rf	250	0.88	0.09	0.85	0.14	0.95	0.11	0.89	0.07	0.78	0.15
rf	500	0.88	0.09	0.85	0.14	0.95	0.11	0.89	0.07	0.78	0.15
knn	1	0.60	0.10	0.60	0.11	0.50	0.25	0.53	0.19	0.20	0.21
knn	5	0.80	0.07	0.85	0.14	0.80	0.27	0.78	0.11	0.66	0.11
knn	9	0.90	0.06	0.88	0.11	0.95	0.11	0.90	0.05	0.82	0.10
nn		0.88	0.13	0.91	0.12	0.85	0.22	0.86	0.15	0.77	0.23
BDC 1 vs R 1		**Accuracy**		**Precision**		**Recall**		**F1 Score**		**MCC**	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.90	0.10	0.91	0.12	0.90	0.14	0.90	0.11	0.81	0.21
rf	250	0.90	0.10	0.91	0.12	0.90	0.14	0.90	0.11	0.81	0.21
rf	500	0.90	0.10	0.91	0.12	0.90	0.14	0.90	0.11	0.81	0.21
knn	1	0.43	0.14	0.43	0.16	0.40	0.14	0.41	0.14	-0.15	0.29
knn	5	0.73	0.10	0.87	0.18	0.60	0.29	0.66	0.18	0.51	0.20
knn	9	0.75	0.09	0.80	0.19	0.75	0.18	0.75	0.08	0.54	0.19
nn		0.85	0.10	0.93	0.15	0.80	0.21	0.84	0.12	0.74	0.18
BDC 1 vs R 3		**Accuracy**		**Precision**		**Recall**		**F1 Score**		**MCC**	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.83	0.14	0.87	0.19	0.80	0.11	0.83	0.13	0.66	0.29
rf	250	0.85	0.10	0.88	0.16	0.85	0.14	0.85	0.09	0.73	0.20
rf	500	0.85	0.10	0.88	0.16	0.85	0.14	0.85	0.09	0.73	0.20
knn	1	0.43	0.17	0.43	0.18	0.40	0.14	0.41	0.15	-0.15	0.34
knn	5	0.68	0.19	0.75	0.26	0.65	0.14	0.68	0.15	0.37	0.40
knn	9	0.63	0.15	0.66	0.23	0.70	0.21	0.65	0.12	0.29	0.33
nn		0.83	0.11	0.87	0.18	0.85	0.22	0.82	0.12	0.70	0.19
BDC 1 vs R 4		**Accuracy**		**Precision**		**Recall**		**F1 Score**		**MCC**	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.85	0.10	0.92	0.11	0.80	0.27	0.82	0.15	0.74	0.18
rf	250	0.83	0.14	0.85	0.14	0.80	0.27	0.80	0.18	0.68	0.28
rf	500	0.88	0.09	0.92	0.11	0.85	0.22	0.86	0.12	0.78	0.15
knn	1	0.58	0.14	0.73	0.28	0.45	0.21	0.50	0.15	0.20	0.32
knn	5	0.70	0.14	0.75	0.43	0.45	0.33	0.53	0.34	0.45	0.29
knn	9	0.58	0.07	0.53	0.51	0.20	0.21	0.27	0.26	0.20	0.19
nn		0.83	0.11	0.89	0.15	0.80	0.27	0.80	0.14	0.70	0.19

Table 10: Result of Day-pair classification using Non-gene physiological dataset

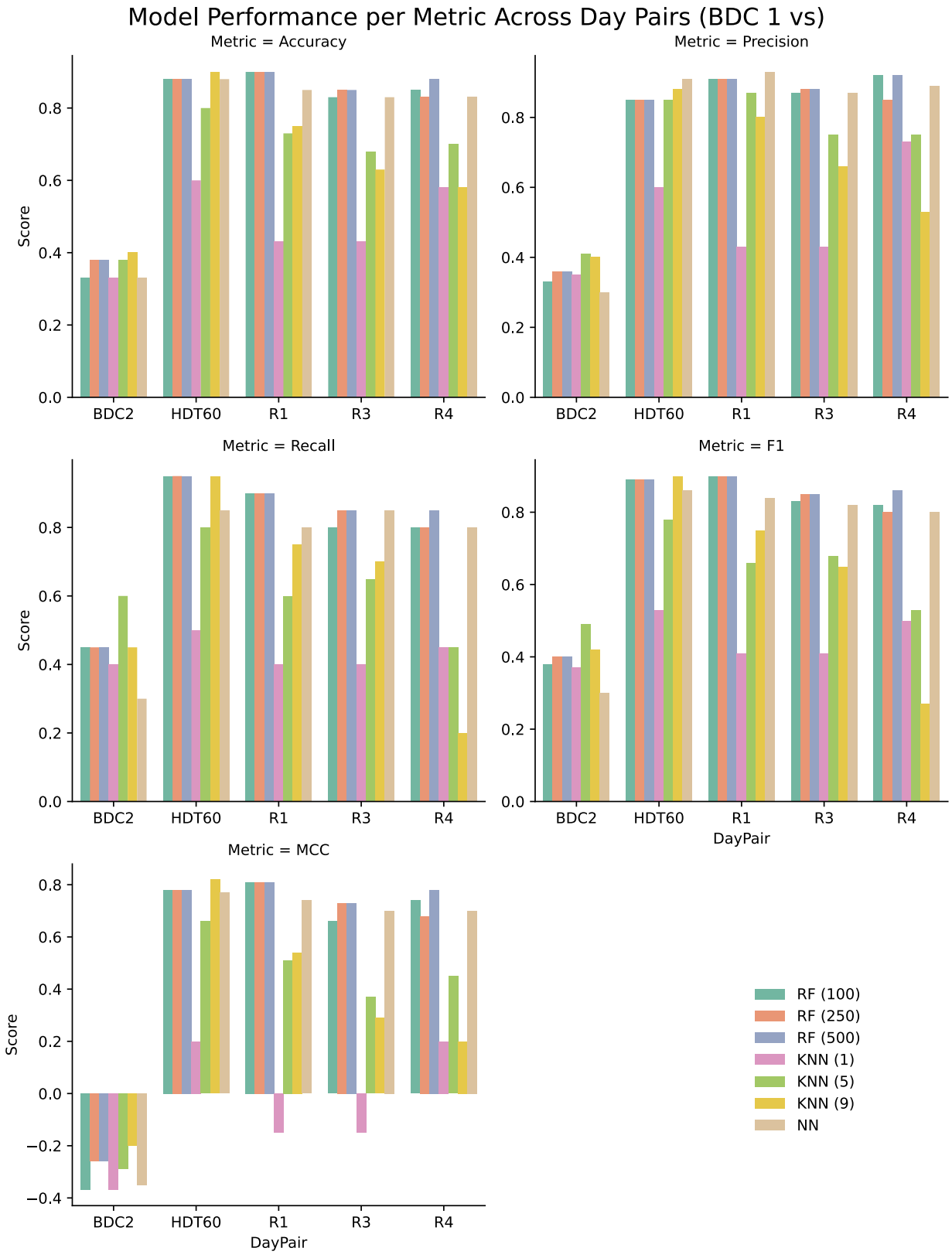


Figure 8: Bar graphs showing the performance metric of Day-pair classification using Non-gene Feature Dataset

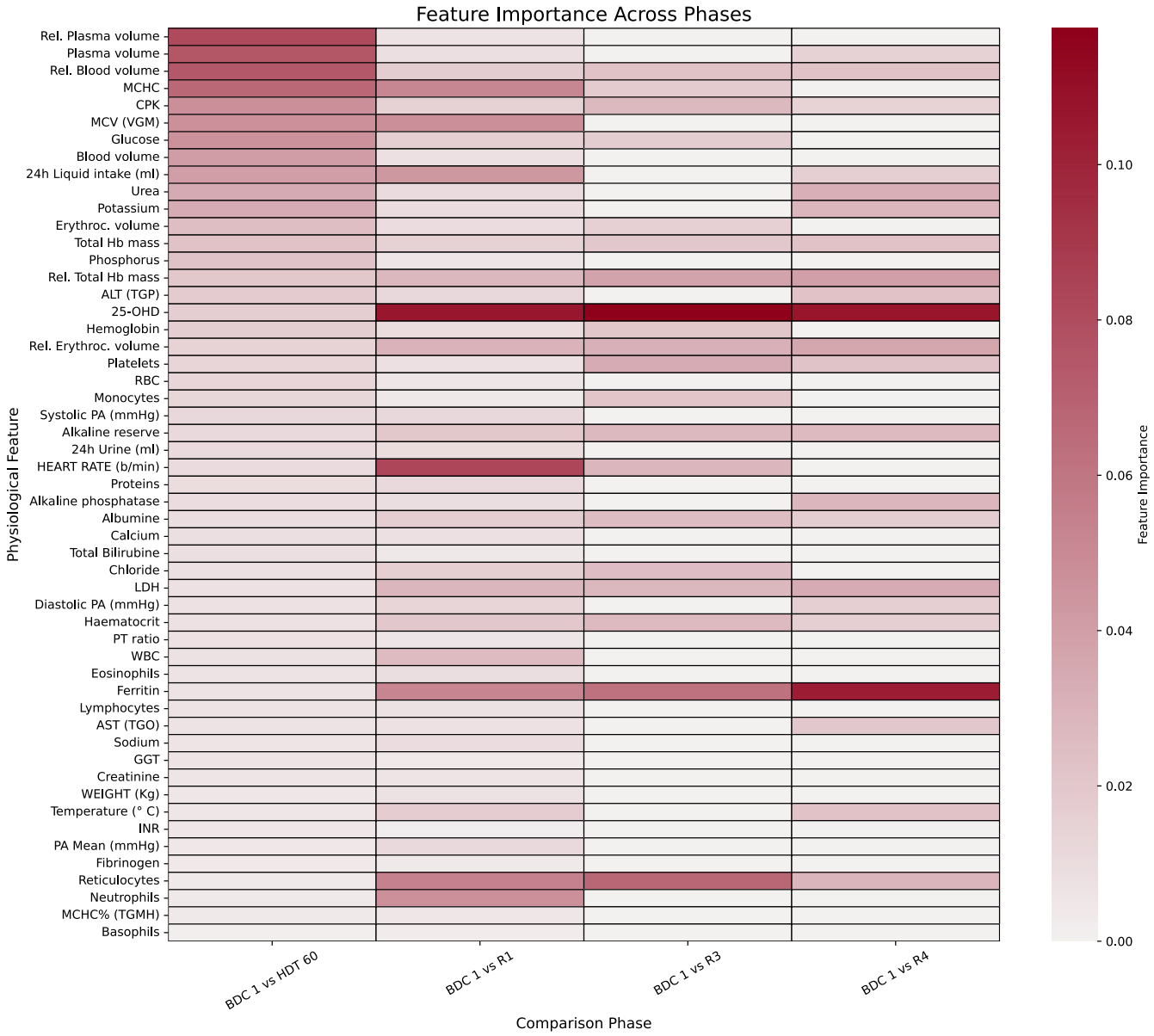


Figure 9: Heatmap showing Feature Importance of Non-Gene features in day-pair classification

Following the quantitative evaluation, we examined feature importance for each day-pair classification with high classification performance. The heatmap highlights features ranked as most discriminative for each day-pair classification consistently. These top features were then analyzed in relation to their biological roles, with specific attention paid to those known to be influenced by inactivity, fluid shift induced by microgravity, and deconditioning which provides insights into the biological processes most affected by head down tilt bed rest and recovery.

For the classification of BDC1 vs HDT 60 using random forest, we achieved an accuracy of ~87% with the feature importance displayed in the heatmap (Figure 9). Relative plasma volume and plasma volume are the top ranked features unique to only this day-pair classification, suggesting a noticeable change in the plasma volume from the baseline to after bed rest. Biologically, plasma volume decreases when a person is subjected to 6-degree head down tilt bed rest [63]. When returning to ambulation, we observed the plasma volume is restored to baseline values. Blood volume and MCHC is influenced directly by the plasma volume, thus the fluid shift influences the entire blood volume homeostasis [91]. An initial change of plasma volume is observed in astronauts while in space [99]. Fluid intake listed among the top features influences plasma volume. The liquid intake is transferred into blood plasma to restore plasma volume to baseline immediately after ending the bed rest [21].

The most consistently important features identified across all classification tasks (BDC1 vs R1, R3, and R4) was 25-hydroxyvitamin D (25-OHD), the primary circulating form of vitamin D. Its high discriminative power suggests that vitamin D metabolism undergoes significant modulation during bed rest and subsequent reambulation. Vitamin D changes are in line with existing literature reporting vitamin D deficiency in astronauts during and after spaceflight, despite controlled dietary supplementation [81]. Furthermore, vitamin D is closely linked to calcium homeostasis, bone remodeling, and immune regulation – all of which are significantly altered in bed rest [5]. During reambulation, vitamin D levels may remain low due to depleted stores, altered metabolism, and increased physiological demand for skeletal reconditioning after an extended period of bed rest [5]. The prominence of 25-OHD in our classification models supports its role as a sensitive biomarker reflecting both bed rest induced musculoskeletal changes and the post bed rest recovery trajectory.

Ferritin also emerged as an important feature across the BDC1 vs R1, R3, and R4 classification tasks, indicating its utility in distinguishing between baseline and recovery phases. Biologically, ferritin serves as the primary intracellular iron storage protein and is a marker of iron status as well as systemic inflammation [56]. In our analysis, ferritin levels were consistently lower in the recovery phases (R1–R4) compared to BDC1, suggesting a potential depletion or redistribution of iron stores during head down tilt and the subsequent reambulation period. While some subjects showed a transient increase in ferritin during the head-down tilt phase – possibly due to acute

inflammatory or hemodynamic shifts [56] – most exhibited a persistent reduction in ferritin levels, aligning with previous findings in space anemia literature [9]. Low ferritin post-bed rest may reflect increased erythropoietic activity, altered iron metabolism, or dilutional effects linked to plasma volume expansion upon return to walking [9]. The prominence of ferritin as a classifier reinforces its role as a sensitive biomarker of physiological stress and iron-demanding recovery processes, particularly in the hematopoietic system.

Reticulocytes, the immature precursors of red blood cells, also ranked among the most important features in distinguishing BDC1 from recovery phases (R1, R3, and R4). Their prominence in feature importance highlights the dynamic changes in erythropoietic activity previously reported during the HDT bed rest phase and at recovery [9]. Notably, our data indicate a dynamic change of reticulocyte concentration during HDT and early R periods, as well as the return to baseline levels at 30 days of reambulation. The transient stimulation of red blood cell production, likely driven by the body's attempt to restore hematologic equilibrium following HDT-induced changes such as plasma volume reduction, hemoconcentration, and red blood cell mass reduction [9]. The change of reticulocyte concentration may reflect the completion of compensatory erythropoiesis, marking a return to homeostasis occurring in the bone marrow. The importance of reticulocytes as a classification feature may stem from subtle shifts in their dynamics or regulation persisting throughout the recovery phase, even when absolute values approximate baseline.

Using Gene transformation Datasets

To investigate whether gene expression profiles reflect distinct physiological states across the experimental timeline, we applied classification models to gene datasets processed through three different normalization techniques log2 transformation**,** DESeq2 normalization, and CLR (centered log-ratio) transformation**.** These models were trained to differentiate BDC1 from subsequent phases using binary classification to test the same two hypotheses from a transcriptomic perspective:

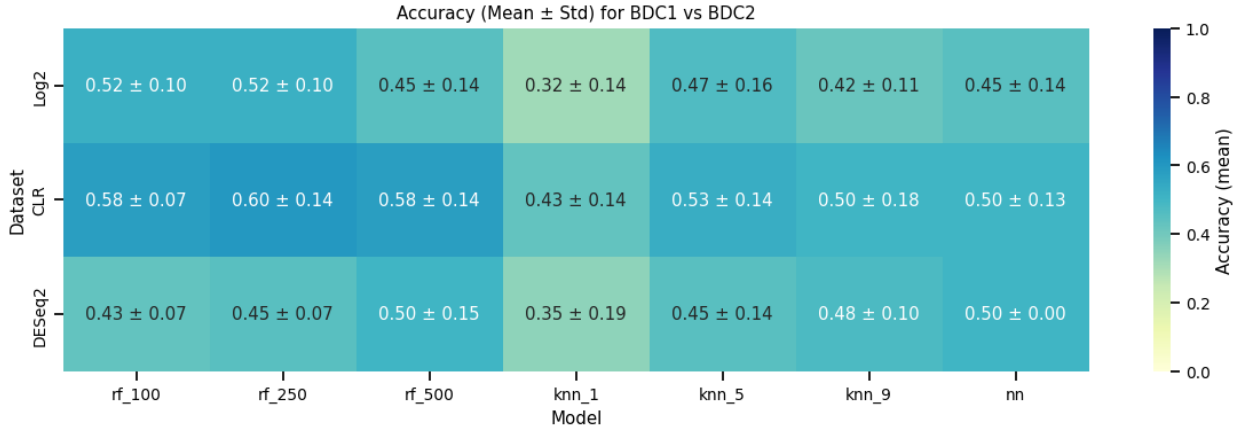
- **Hypothesis H2**: *BDC1 is significantly different from the out-of-phase timepoints (HDT60, R1, and R3).*
  Result: Not Rejected.

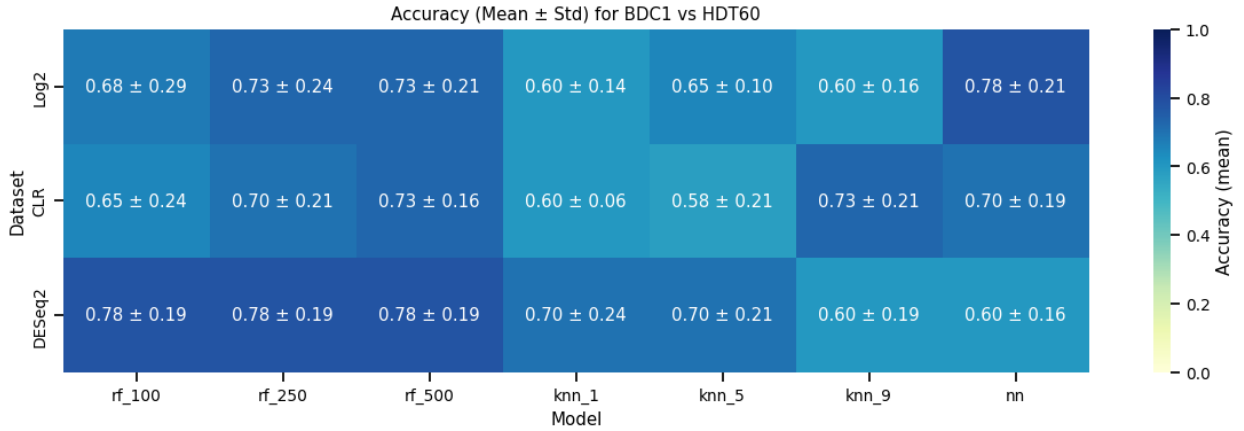
  Across all three transformed gene datasets, the classification models consistently achieved high performance in separating BDC1 from HDT60 and the early recovery phases (R1, R3). This indicates that transcriptional activity was significantly altered during bed rest and did not return to baseline immediately during recovery. These findings suggest that gene-level regulation is sensitive to both the stress of inactivity and the physiological demands of recovery.

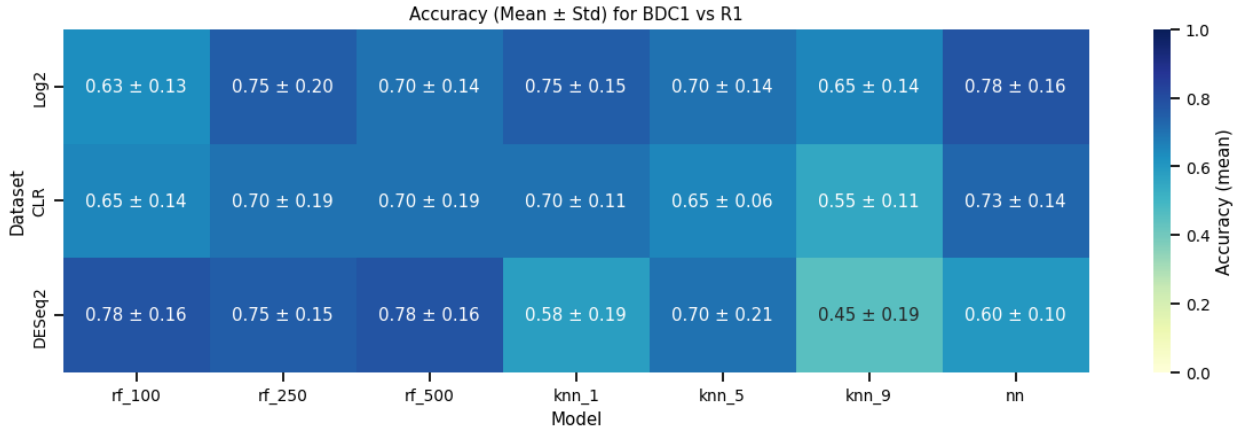
- **Hypothesis H3**: *By R4, participants' gene expression returns to a baseline-like state.*
  Result: Rejected.

  Contrary to expectations, gene expression at R4 remained distinguishable from BDC1 across all transformation methods and classification models. This implies that transcriptional recovery lags behind physiological recovery, or that long-term shifts in gene regulation persist even after the observable recovery period. These lingering transcriptomic differences may point to underlying biological processes that remain active beyond visible physiological stabilization.

All detailed performance metrics including accuracy, precision, recall, F1-score, and MCC, along with their corresponding mean and standard deviation for every day-pair classification across all models and datasets are provided in Appendix II: Classification into phases. To visually highlight classification performance across the three gene transformation datasets, the mean accuracy values for each phase comparison are summarized in the heatmap (Figure 10). This visualization clearly shows that BDC1 vs BDC2 yielded poor separability, while comparisons between BDC1 and later phases (HDT60, R1, R3, R4) resulted in significantly higher classification accuracy, indicating stronger transcriptomic shifts.


Accuracy (Mean ± Std) for BDC1 vs BDC2
Dataset
Log2
CLR
DESeq2
0.52 ± 0.10
0.52 ± 0.10
0.45 ± 0.14
0.32 ± 0.14
0.47 ± 0.16
0.42 ± 0.11
0.45 ± 0.14
0.58 ± 0.07
0.60 ± 0.14
0.58 ± 0.14
0.43 ± 0.14
0.53 ± 0.14
0.50 ± 0.18
0.50 ± 0.13
0.43 ± 0.07
0.45 ± 0.07
0.50 ± 0.15
0.35 ± 0.19
0.45 ± 0.14
0.48 ± 0.10
0.50 ± 0.00
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
Accuracy (mean)
1.0
0.8
0.6
0.4
0.2
0.0


Accuracy (Mean ± Std) for BDC1 vs HDT60
Dataset
Log2
CLR
DESeq2
0.68 ± 0.29
0.73 ± 0.24
0.73 ± 0.21
0.60 ± 0.14
0.65 ± 0.10
0.60 ± 0.16
0.78 ± 0.21
0.65 ± 0.24
0.70 ± 0.21
0.73 ± 0.16
0.60 ± 0.06
0.58 ± 0.21
0.73 ± 0.21
0.70 ± 0.19
0.78 ± 0.19
0.78 ± 0.19
0.78 ± 0.19
0.70 ± 0.24
0.70 ± 0.21
0.60 ± 0.19
0.60 ± 0.16
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
Accuracy (mean)
1.0
0.8
0.6
0.4
0.2
0.0


Accuracy (Mean ± Std) for BDC1 vs R1
Dataset
Log2
CLR
DESeq2
0.63 ± 0.13
0.75 ± 0.20
0.70 ± 0.14
0.75 ± 0.15
0.70 ± 0.14
0.65 ± 0.14
0.78 ± 0.16
0.65 ± 0.14
0.70 ± 0.19
0.70 ± 0.19
0.70 ± 0.11
0.65 ± 0.06
0.55 ± 0.11
0.73 ± 0.14
0.78 ± 0.16
0.75 ± 0.15
0.78 ± 0.16
0.58 ± 0.19
0.70 ± 0.21
0.45 ± 0.19
0.60 ± 0.10
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
Accuracy (mean)
1.0
0.8
0.6
0.4
0.2
0.0

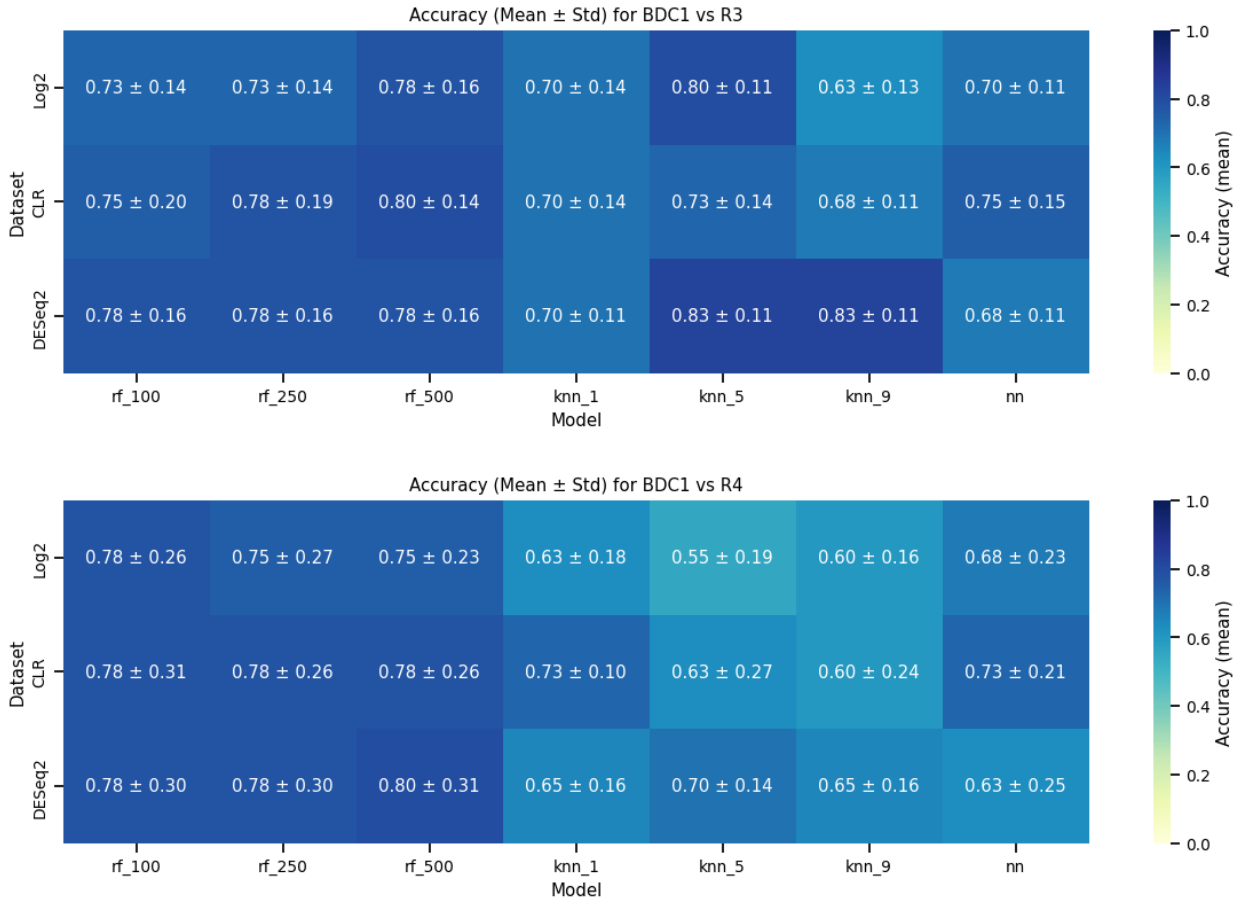


Figure 10: Heatmap of mean accuracy across different Gene-transformation datasets

To further explore the biological significance behind these separations, an UpSet plot is presented, capturing the intersections of highly informative genes across datasets and day-pair comparisons with strong classification performance. This allows for identification of common and phase-specific gene signatures, offering insight into the transcriptional pathways affected during bed rest and the recovery process. These features and their functional relevance are examined in detail in the discussion that follows.

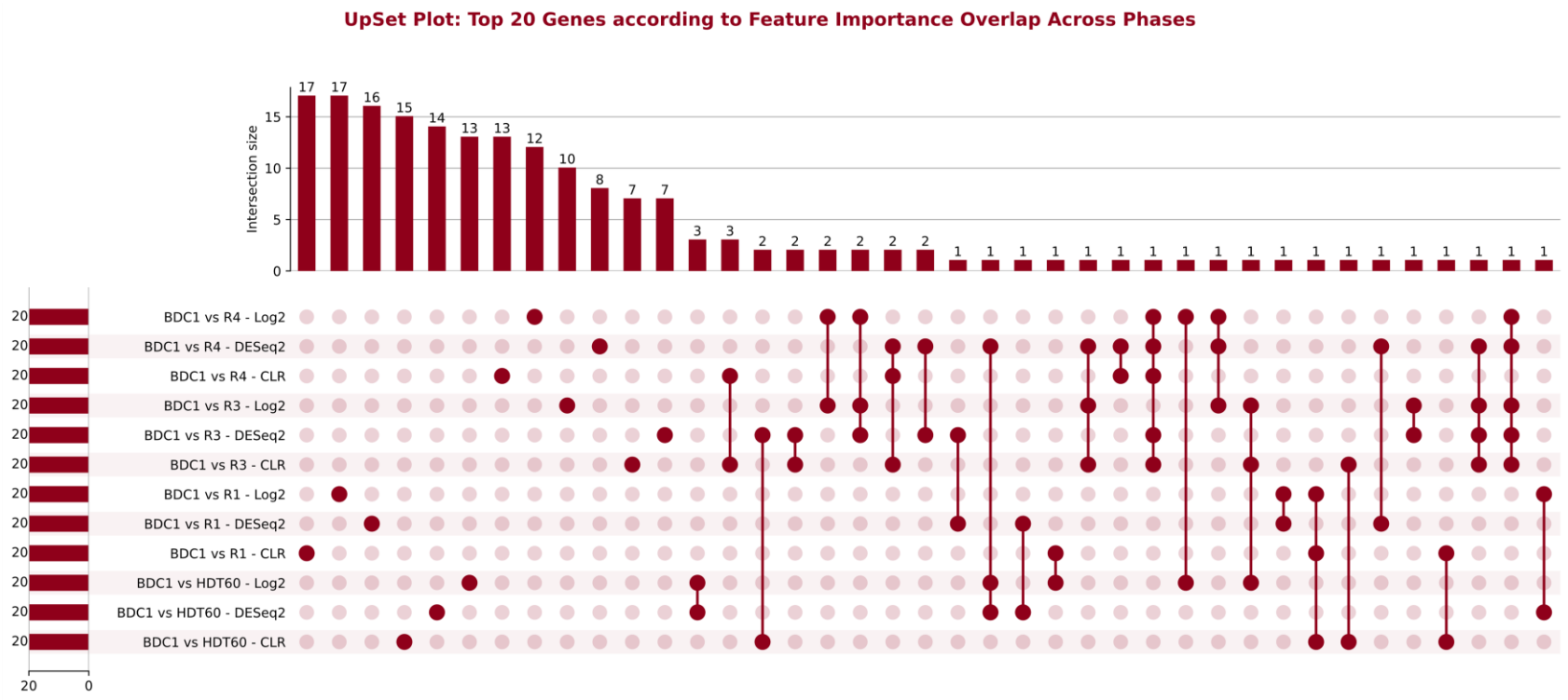


Figure 11: Upset plot showing the top 20 genes identified by Random Forest feature importance across multiple classification tasks comparing baseline (BDC1) with different phases (HDT60, R1, R3, R4) using three gene transformation methods (Log2, DESeq2, CLR). Each intersection represents a unique set of genes shared between specific classification tasks. The bar plot on the left indicates the number of times each gene appeared in the top 20 across tasks, while the bottom bar plot shows the total number of top genes identified per classification task. This visualization highlights genes consistently important across transformations and timepoints, indicating potential biological relevance to microgravity adaptation and recovery.

Table 11: Some common genes from the upset plot are listed in the table below.

Criteria 1 **HDT 60 ∩ R1/R3/R4**	**Criteria 2** **R1/R3 ∩ R4**
ENSG00000106993, ENSG00000171155, ENSG00000110090, ENSG00000110955, ENSG00000158488, ENSG00000267576, ENSG00000157445, ENSG00000221792, ENSG00000129055, ENSG00000038219, ENSG00000270137	ENSG00000179611, ENSG00000229180, ENSG00000232149, ENSG00000164889, ENSG00000178950, ENSG00000176973, ENSG00000257342, ENSG00000112309, ENSG00000271980, ENSG00000105556, ENSG00000245970, ENSG00000262160, ENSG00000212802, ENSG00000225798, ENSG00000269246

Common Genes and Their Biological Significance in Day Pair Classification

This section describes the common top genes identified through feature importance analysis across various gene transformation datasets in the day pair classification. These genes are categorized based on two distinct criteria to gain insight into their relevance during bed rest (HDT60) and recovery (R1, R3, R4) relative to baseline (BDC1).

Criteria for Gene Categorization

We categorize the commonly observed top genes based on their feature importance across all gene transformation datasets using two criteria:

Criteria 1: Persistent Differences from Baseline to Bed Rest and Post-Bed Rest. The objective was to identify common top genes in the intersection of "BDC1 vs HDT60" and the union of "BDC1 vs R1", "BDC1 vs R3", and "BDC1 vs R4". The resulting set of genes indicates persistent differences from baseline (BDC1) during bed rest (HDT60) as well as throughout the recovery phases (R1, R3, R4). Analyzing these genes provides insight into those that exhibit altered expression or importance across all phases of the study (bed rest and recovery) compared to their baseline values.

Intersection Logic: (BDC1 vs HDT60) ∩ ((BDC1 vs R1) ∪ (BDC1 vs R3) ∪ (BDC1 vs R4))

Criteria 2: Dynamics of Recovery Relative to Baseline After HDT60. This criterion identifies common top genes within the union of "BDC1 vs R1" and "BDC1 vs R3",

which then intersect with "BDC1 vs R4". This set of genes highlights the dynamics of recovery relative to baseline following the HDT60 (bed rest) period. It aims to capture genes that are consistently important in later recovery phases (R4) but may also be important in earlier recovery phases (R1 or R3).

Intersection Logic: ((BDC1 vs R1) ∪ (BDC1 vs R3)) ∩ (BDC1 vs R4)

Identified Common Genes

The details of the common genes identified for both criteria are presented in Table 12 for Criteria 1 and Table 13for Criteria 2. It is important to note that no common genes were found to satisfy both Criteria 1 and Criteria 2 simultaneously, indicating distinct sets of genes driving the observed changes under these two different perspectives.

Table 12: Gene ENS # and GO terms for Criteria 1

Criteria 1		
Gene ENSG #	Gene Name, (Gene Symbol)	GO term (biological process) (from GeneCards)
ENSG00000106993	Cell Division cycle 37 Like 1, HSP90 Cochaperone (CDC37L1)	Involved in: protein folding, protein stabilization
ENSG00000171155	C1GALT1 Specific Chaperone 1 (C1GALT1C1)	involved in: protein O-linked glycosylation, platelet activation, platelet morphogenesis
ENSG00000110090	Carnitine Palmitoyltransferase 1A (CPT1A)	involved in: response to hypoxia, long-chain fatty acid metabolic process, glucose metabolic process, fatty acid metabolic process, fatty acid beta-oxidation
ENSG00000110955	ATP Synthase F1 Subunit Beta	involved in: angiogenesis, osteoblast differentiation, generation of precursor metabolites

	(ATP5F1B)	and energy, lipid metabolic process, ATP biosynthetic process
ENSG00000158488	CD1e Molecule (CD1E)	involved in: positive regulation of T cell mediated cytotoxicity, adaptive immune response, immune response, endogenous lipid antigen via MHC class Ib, antigen processing and presentation, exogenous lipid antigen via MHC class Ib
ENSG00000267576	Novel transcript, antisense to TSPAN16 RNA Gene (lncRNA)	GO term unavailable
ENSG00000157445	Calcium Voltage-Gated Channel Auxiliary Subunit Alpha2Delta 3 (CACNA2D3)	involved in: calcium ion transmembrane transport, regulation of presynapse organization
ENSG00000221792	MIR1282 MicroRNA 1282 RNA Gene (miRNA)	GO term unavailable
ENSG00000129055	Anaphase Promoting Complex Subunit 13 (ANAPC13)	protein ubiquitination; involved in: regulation of mitotic cell cycle, anaphase-promoting complex-dependent catabolic process, cell division, regulation of meiotic cell cycle
ENSG00000038219	Biorientation Of Chromosomes In Cell Division 1 Like 1 protein coding (BOD1L1)	involved in: DNA damage repair, DNA damage response, replication fork processing
ENSG00000270137	Novel Transcript, intronic to PHF20L1 RNA Gene (lncRNA)	GO term unavailable

Eleven genes, including (n = 2 genes) lncRNAs – long non-coding RNA, (n = 1) miRNA – micro RNA, and (n = 8) protein-coding RNAs identified in Criteria 1 provide compelling insights into the physiological systems affected by bed rest and/or microgravity. These genes, highlighted in Table 12, suggest disruptions in immune, bone, nervous, and metabolic processes and are discussed below.

To begin with, the appearance of CDC37L1 in Criteria 1 suggested changes in protein synthesis and protein homeostasis induced by inactivity, microgravity, or both during bed rest and after (Table 12, Appendix III: Synthetic Data Results). Evidence of altered protein metabolism in immune cells of astronauts in spaceflight exposed to microgravity has been reported in previous studies [58]. Next, C1GALT1C1 from the list Criteria 1, is involved in the folding of O-glycans of the T antigen; a precursor for many glycoproteins [92]. Studies also reported a role for C1GALT1C1 in platelet activation and platelet morphogenesis suggesting altered platelet activity during bed rest. In support for platelet changes, a different 60-day HDT bed rest study described changes in platelets' concentration [82].

Two genes CPT1A and ATP5F1B from Criteria 1 list are involved in glucose and/ or lipid metabolism suggesting a perturbation of energy metabolism. Our results identified blood glucose levels among the top ranked non-gene features supporting changes in overall metabolism during the bed rest and reambulation phases (Figure 9). Researchers studying the cocktail countermeasure also reported disturbances in glucose removal from the bloodstream [59] [60] and lipid oxidation [10], all contributing to energy and metabolism. The reduced expression of ATP5F1B gene, involved in osteoblast differentiation, during HDT60 and initial R is in line with previous reports of bone density loss in microgravity and bed rest studies [6]. The presence of CD1E in Criteria 1 suggested changes in the immune system, as CD1E is involved in the immune response to pathogens and antigen presentation. Immune functions are affected by microgravity and hospital bed rest. In vitro microgravity studies of gene expression in macrophages described opposite effects of short-term microgravity, and of hyper gravity [64]. A microgravity study using the U937 macrophage cell line showed significant decrease in MHC-II (major histocompatibility complex class I which is responsible for presenting antigens to immune cells) levels after 5 days of simulated microgravity [68]. Immune cells change observed in vitro models might not be observed using in vivo model such as HDT bed rest. Nevertheless, the impact of microgravity on immune

cells activity is well established. Lastly, CACNA2D3 from the list Criteria 1 is involved in calcium ion transmembrane transport and regulation of presynapse organization. The response of the nervous system to a change of gravitational force is less understood. Change in nervous cells function and signaling in response to microgravity or bed rest or both have been described. Microgravity has been shown to alter expression of genes involved in the differentiation and regulation of ion channels and simulated microgravity altered neuron-astrocyte communication [66].

Table 13: Gene ENS # and GO terms of Criteria 2

Criteria 2		
Gene ENS #	Gene Name, (Gene Symbol)	GO term (biological process) (from GeneCards)
ENSG00000179611	Diaglycerol Kinase Zeta Pseudogene 1 (DGKZP1)	GO term unavailable
ENSG00000229180	RABGEF1P1 RABGEF1 Pseudogene 1	GO term unavailable
ENSG00000232149	FERP1 FER tyrosine kinase Pseudogene 1	GO term unavailable
ENSG00000164889	SLC4A2 Solute Carrier Family 4 Member 2	monoatomic ion transmembrane transporter activity ; bicarbonate transmembrane transporter activity ; enables: solute: inorganic anion antiporter activity / protein binding / monoatomic anion transmembrane transporter activity
ENSG00000178950	GAK Cyclin G Associated Kinase	protein phosphorylation ; involved in : chromatin remodeling /. receptor-mediated endocytosis. /. endoplasmic reticulum organization. /. Golgi organization

ENSG00000176973	FAM89B Family With Sequence Similarity 89 Member B	involved in: establishment of cell polarity /. positive regulation of cell migration. /. negative regulation of transforming growth factor beta receptor signaling pathway. / negative regulation of SMAD protein signaling transduction
ENSG00000257342	Novel Transcript, Antisense To OS9 lncRNA	GO term unavailable
ENSG00000112309	B3GAT2 Beta-1,3-Glucuronyltransferase 2	involved in: carbohydrate metabolic process. /. protein glycosylation. /. carbohydrate biosynthetic process. /. chondroitin sulfate proteoglycan biosynthetic process
ENSG00000271980	Novel Transcript lncRNA	GO term unavailable
ENSG00000105556	MIER2 MIER Family Member 2	involved in: negative regulation of transcription by RNA polymerase II ; chromatin remodeling
ENSG00000245970	RPL30-AS1 RPL30 Antisense RNA 1 lncRNA	GO term unavailable
ENSG00000262160	Novel Transcript, Antisense To NFATC3 lncRNA	GO term unavailable
ENSG00000212802	RPL15P3 Ribosomal Protein L15 Pseudogene 3	GO term unavailable
ENSG00000225798	novel transcript lncRNA	GO term unavailable
ENSG00000269246	novel transcript lncRNA	GO term unavailable

Criteria 2 identified 15 genes including lncRNAs (n = 6 genes), pseudogenes (n = 4) and protein-coding genes (n = 5). Discussion focused on protein-coding genes as functional information available for pseudogenes and lncRNA is limited.

The gene SLC42A in Criteria 2 encodes the protein Anion Exchanger 2 (AE2), member of the chloride-bicarbonate exchanger AE subfamily from the bicarbonate-transporters SLC4 family [67]. Aside from regulation of cellular pH in various tissues, the AE2 exchangers provide a cytosolic acid reservoir in cells specialized in acid secretion such as gastric parietal cells and osteoclasts [69]. Bone resorption has been shown to increase after space flight [81]. Our findings on altered expression of SLC42A could signal an increase in EA2 protein to support the enhanced activity of osteoclasts in the microgravity bed rest analogs and spaceflight.

GAK gene from Criteria 2 is involved in protein phosphorylation, as well as chromatin remodeling and receptor-mediated endocytosis. A proteomic study has been conducted on human primary T-lymphocytes signaling cascade after exposure to short-term simulated microgravity (20s parabolic flight, 5m clinorotation and 60m clinorotation). The analysis showed an increase in acetylated histone H3 and variations in different surface signaling proteins in non-activated and activated lymphocytes following the 5min clinorotation compared to baseline cell samples and 1g controls [65]. Our findings would support the idea that immune cell signaling and interaction with chromatin were perturbed during the 60-day HDT bed rest and continued through reambulation.

FAM89B gene identified under Criteria 2 is of interest due to its consistent importance in classification tasks involving the reambulation phase. A previous study performed DNA microarray analysis on yeast cell cultures growing in a low-shear modeled microgravity. They reported that impaired establishment of cell polarity is due to an increase or decrease ranging from 1 to 5.3 fold in the expression of multiple genes responsible in cell polarity establishment [71]. The changing levels of FAM89B in our analysis would support the idea of analogous changes happening in human cells in response to microgravity. FAM89B is also involved in cell migration – this process is reported to be significantly inhibited in primitive bone marrow CD34+ cells exposed to microgravity (hemopoietic stem cells), possibly contributing to hemostasis disturbances described in astronauts [72]. Lastly, FAM89B is involved in negative regulation of transforming growth factor beta – SMAD receptor signaling pathway, both playing roles in

immunity, with a primary function of regulation of CD4(+) T cells to suppress or enhance the body's immune response [73].

B3GAT2 gene, identified in Criteria 2, is involved in protein glycosylation, carbohydrate biosynthetic processes and chondroitin sulfate proteoglycans (CSPGs) biosynthetic processes. Thus, change during the reambulation phase compared to baseline suggests changes in cell membrane protein and glyco-protein expression contributing to cell adhesion, migration, signaling, and other similar processes [74] [75].

We also observed MIER2 gene in Criteria 2, a gene involved in the negative regulation of transcription by RNA polymerase II and chromatin remodeling. Those two processes enable MIER2 to decrease transcription of a variety of genes. Such a mechanism could have contributed to the temporally altered gene expression profile observed in the Cocktail study [22].

Additional analyses after application of dimensionality reduction as well as using a list of genes limited to protein coding genes could give us more information on important genes for classification into day-pairs.

Results - Task 3: Synthetic Gene and Non-gene Data

To evaluate the sensitivity of our models to intervention signals of varying strengths and data availability, we generated synthetic datasets for both gene expression and physiological features using a parametric perturbation-based simulation. In this approach, a perturbation parameter a was used to control the intensity of deviation from a reference sample, with synthetic control and intervention (cocktail) groups created through random variation and mean-shifted noise, respectively.

This framework allowed us to simulate data under varying signal strengths (values of a) and sample sizes (n = 20, 200, 2000, 20000). For each configuration, we evaluated classification performance using multiple models: Random Forests (with 50, 100, and 150 estimators), K-Nearest Neighbors (with k = 1, 5, 9) and a feedforward Neural Network. The primary evaluation metric was classification accuracy, with additional metrics presented in Appendix III: Synthetic Data Results.

Motivation

The core motivation for this simulation was to test whether the intervention signal (if any) was learnable under different levels of sample size and magnitude of perturbation values. This is especially important in biological contexts where collecting large amounts of high-quality labeled data can be costly or infeasible. Our simulation enabled us to explore whether observed effects would remain classifiable under weaker signals or limited data sizes.

Key Observations

- High perturbation values consistently resulted in high classification accuracy, even with low sample sizes (n = 20). This suggests that when the intervention effect is strong, it is clearly distinguishable even with smaller datasets.
- At lower perturbation, classification accuracy dropped across all models and datasets, and this decline persisted even as the sample size increased. In other words, merely increasing the number of synthetic samples did not improve the model's ability to detect the signal when the perturbation effect was small.

- This result demonstrates that signal strength (effect size) can be a more critical factor than sample size for successful classification in this setting.
- Both gene-based and physiological feature-based datasets followed this trend. While gene data tended to show more dramatic changes in accuracy with varying a, physiological features displayed smoother but still consistent declines.

These findings suggests that our simulation strategy can be a valid approach to approximate how classification models would perform under varying dataset sizes and intervention signal strengths.

In the real dataset, we observed relatively low sample sizes per group (control vs. cocktail). The classification performance was not uniformly strong, indicating the intervention did not induce a consistent or uniform effect across all subjects**.** This is aligned with the patterns observed in our synthetic simulations at lower a value. The absence of a robust and consistent signal – both real and simulated – supports the conclusion that the intervention may not have led to a strong enough physiological or genomic response to be reliably classifiable.

As a result, we reject Hypothesis 1, which assumed a uniform and classifiable effect of the intervention. We argue that this rejection is not due to a methodological limitation or insufficient sample size alone, but rather because the underlying intervention effect appears to be weak.

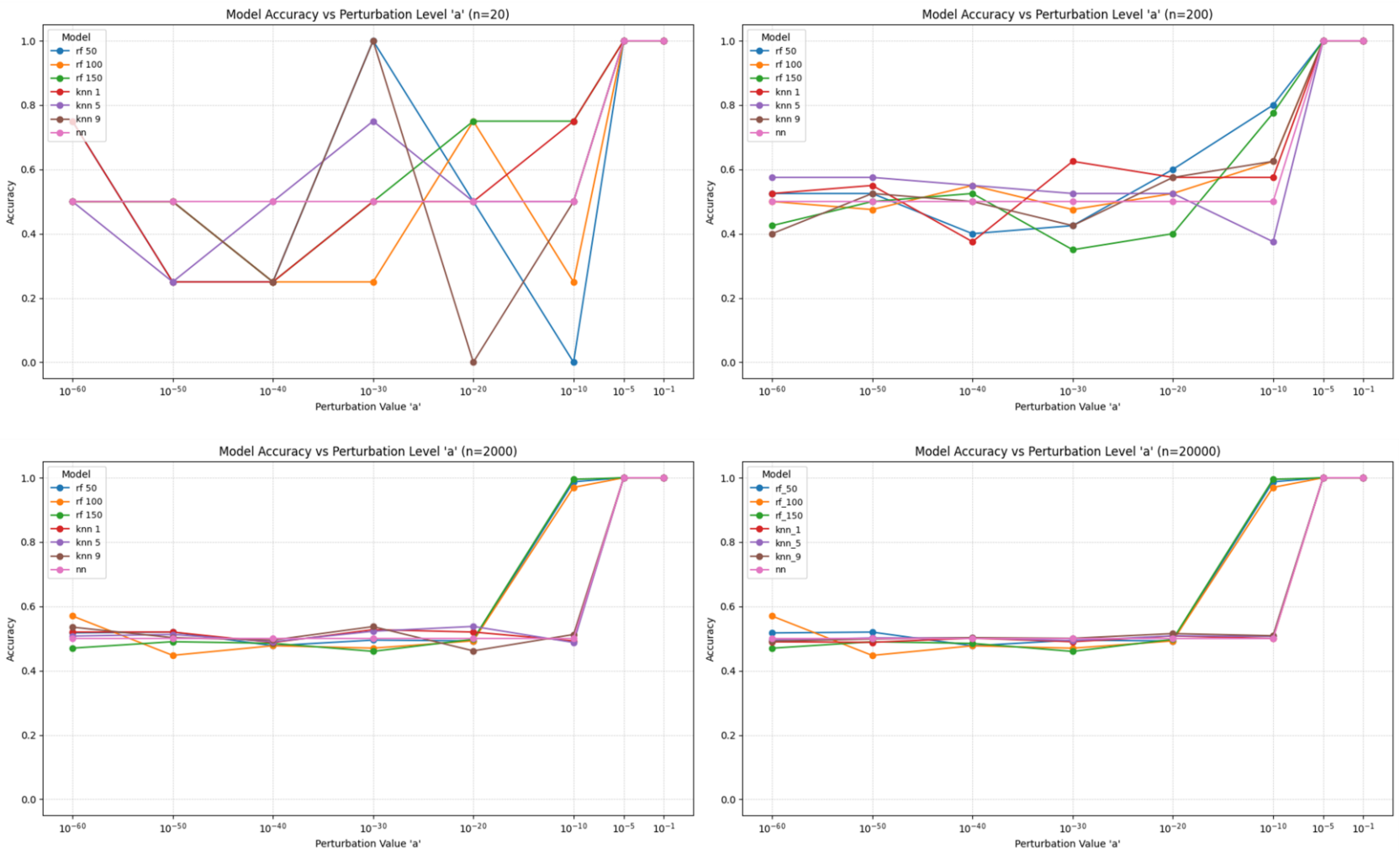


Figure 12: Synthetic Gene Data accuracy for multiple sample size and perturbation values

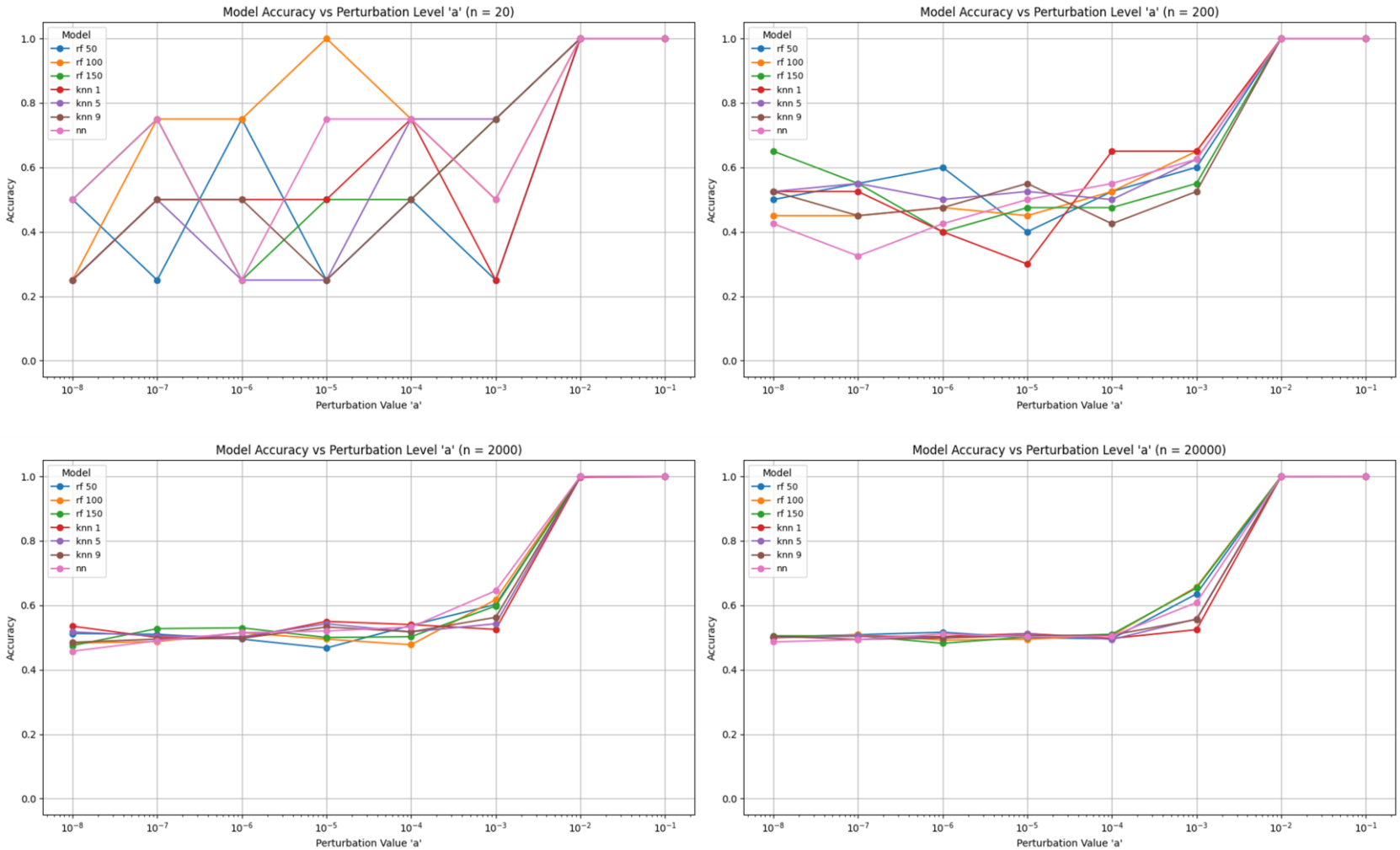


Figure 13: Synthetic Non-Gene Feature Data accuracy for multiple sample size and perturbation values

Limitations and Future Work

While this study provides important insights into the physiological and transcriptomic responses to prolonged bed rest and nutritional intervention, several limitations must be acknowledged to contextualize the findings and guide future improvements. These limitations arise from both experimental constraints – such as the availability of preprocessed gene expression data and the design of the nutritional countermeasure – and methodological decisions made during preprocessing, feature selection, and model evaluation. Recognizing these challenges is essential for refining future analyses, improving model interpretability, and enhancing the biological relevance of the results. In the following subsections, we outline key limitations encountered during data processing and modeling and propose new directions for future work to overcome them.

Data Preprocessing Choices and Their Constraints

A critical component of this study involved preparing the data in a form suitable for machine learning, which required several preprocessing decisions. While these choices were carefully made in an attempt to preserve biological meaning and ensure model compatibility, they also introduced certain constraints and assumptions that could have influenced the downstream results. This section discusses the rationale behind the preprocessing strategies, their limitations, and how they might be improved in future analyses. The first preprocessing decision involved the exclusion of features from the nutrition category. Although nutrition was carefully controlled and designed to be balanced across all participants, the meals provided to subjects varied daily and among individuals. Including them would have introduced noise into the model, potentially masking true biological signals and confounding classification performance. Since our study aimed to assess biological responses to bed rest and nutritional intervention rather than daily food variation, we removed these features entirely from the classification input.

For certain features, missing values were imputed using the value from the nearest timepoint in the respective phase rather than global statistics (mean or median). This approach was selected because many biological parameters exhibit strong temporal dependence and intra-subject continuity. Using a temporally adjacent value preserved the longitudinal structure of the data and prevented unrealistic flat imputation across all individuals.

Limitations of DESeq2 and Motivation for Alternative Transformations

Regarding the gene expression data, our analysis was constrained by the form in which it was provided. The DESeq2-normalized gene expression values were received from the primary investigator. These values are widely used in transcriptomic studies, but they represent a transformation of raw read counts – not raw data. As such, our preprocessing options were limited. We employed two alternative transformations: log2 transformation and centered log-ratio (CLR) transformation.

- The log2 transformation is commonly used to stabilize variance and normalize highly skewed expression distributions, making the data more suitable for models like KNN and Random Forest [42].
- The CLR transformation, on the other hand, is recommended for compositional gene expression data, where relative rather than absolute abundance is informative. Since transcriptomics is inherently compositional (total read counts are constrained), CLR offers a biologically meaningful view that allows comparison of genes relative to the geometric mean of all expressed genes in a sample.

Both transformations were applied in parallel and used for model training and feature importance evaluation. These approaches slightly enhanced classification performance across days and groups which in turn improved model interpretability. Future work could investigate additional compositional methods such as ALR (additive log-ratio) [37] or ILR (isometric log-ratio) [109] in more depth to compare their usefulness and interpretability.

Stable Day Classification and Cross-Dataset Consistency

An important observation was that the classification accuracy across the four different datasets – physiological features and gene-normalized sets using DESeq2, log2, and CLR was consistently high when distinguishing between early and late phases of the study (e.g., BDC vs. R). This strengthens the reliability of our labels and the underlying experimental design, confirming that the temporal component (days) was a strong source of signal. However, the classification accuracy was not sufficient for all countermeasure classification or all datasets, especially in those where the intervention effect may have been weak or delayed. In future studies, longitudinal modeling

approaches (e.g., time-series deep learning or mixed-effects models) could offer a more nuanced view of temporal dynamics.

Dimensionality Reduction: Alternatives Explored and Final Choice

While automated dimensionality reduction techniques such as PCA, t-SNE, and kernel PCA were not implemented in the current analysis, they represent valuable avenues for future exploration [40] [90] [94]. Our primary aim in this study was to preserve biological interpretability, which required us to retain all transcript-level gene data and focus on domain-relevant non-gene features such as vital signs, blood parameters, plasma volume, and liquid balance. Rather than applying unsupervised techniques that may obscure biological signals in favor of global variance patterns, we adopted a knowledge-driven feature selection approach. This allowed us to emphasize variables of known physiological relevance while maintaining the richness of the full transcriptomic dataset. Future work could benefit from hybrid strategies that combine unsupervised dimensionality reduction with supervised relevance scoring to balance interpretability and model efficiency.

Model Choice and Hyperparameter Justification

Model selection in this study was intentionally designed to prioritize interpretability and biological insight over performance optimization. Rather than fine-tuning for maximum accuracy, we focused on evaluating whether separability between experimental groups – such as cocktail vs control or different timepoints – was strong enough to be detected using standard, widely accepted machine learning models [24] [28] [31]. This strategy aligns with the study's core objective: to explore systemic effects of the countermeasure rather than to build a highly optimized predictive tool.

We selected three families of models that are commonly used in biomedical and omics-related classification tasks. The K-Nearest Neighbors (kNN) classifier was used as a simple, non-parametric baseline model that assumes local homogeneity and makes no prior assumptions about data distribution [28]. Its distance-based classification is often effective in biological contexts where individuals with similar physiological or transcriptomic profiles are expected to cluster together. Random Forest (RF) was chosen due to its ensemble-based stability, ability to handle both categorical and continuous features, and natural mechanism for feature importance extraction [24]. We used 50, 100, and 150 estimators to assess consistency in classification across ensemble

sizes without the need for extensive hyperparameter tuning. Lastly, Neural Networks (NNs) were employed to explore whether non-linear patterns and cross-feature interactions could be captured even in relatively low-sample, high-dimensional contexts like ours [31]. A feedforward architecture was implemented with default layer sizes, dropout regularization, and ReLU activation.

These models are broadly validated in similar domains – particularly in biomedical studies involving omics data integration, physiological state monitoring, and treatment outcome classification – making them appropriate choices for initial experimentation. Their use provided insights into the structural differences between groups in the dataset, even without sophisticated tuning.

For future work, more sophisticated modeling strategies could be explored to further validate and potentially enhance classification performance. These include automated machine learning (AutoML) platforms, which systematically search through model types and hyperparameter spaces to find the best configurations. Additionally, nested cross-validation could be implemented for rigorous hyperparameter tuning while avoiding overfitting [26]. More complex neural architectures, such as convolutional neural networks (CNNs) for structured gene data or attention-based models, could also be tested to assess whether deeper models uncover subtler patterns missed by shallow networks. Ultimately, these advanced methods may offer improved performance and new biological insights, but our current approach demonstrates measurable group differences with basic models sufficient to test our hypotheses.

Synthetic Data: Sample Size vs. Perturbation Signal

Our use of generative modeling via perturbation-based simulation enabled a controlled analysis of how well models could distinguish intervention groups under varying signal strengths and sample sizes (from 20 to 20,000). A key finding was that increasing sample size does not improve classification performance when the perturbation (signal) is low. This suggests that the intensity of the intervention effect can be more critical than the quantity of available data. This validates our results even in the presence of small real-world sample sizes and supports the notion that our inability to classify the cocktail vs. control group robustly was not a methodological artifact but a true biological observation. This insight supports the rejection of Hypothesis 1.

In future work, other data augmentation techniques (such as generative adversarial networks, variational autoencoders, or bootstrapping with added biological noise) could complement perturbation modeling [84].

Explainability and Model Interpretability

While we used traditional metrics like accuracy and cross-validation to evaluate models, feature importance and mechanistic interpretability were not exhaustively explored in this phase. In future work, we plan to incorporate alternative explainable techniques like SHAP (SHapley Additive exPlanations) value [96] analysis to offer deeper insights into which features (genes or physiological signals) drive classification. This will not only enhance transparency but may also further help biological interpretability.

Experimental Design and Repeated Measures

Finally, our experimental design was based on repeated measures per subject, a structure that inherently captures within-subject variability and increases power when used with appropriate modeling strategies (e.g., mixed-effects models). However, our current models treated each row independently. Future version of the experiment could incorporate subject-level random effects or time-aware modeling to make full use of the longitudinal structure.

In summary, while the current study effectively demonstrates the feasibility of detecting physiological and transcriptomic signatures of prolonged inactivity and intervention using machine learning, several methodological and experimental constraints shaped the scope of our findings. These included preprocessing limitations, simplified modeling strategies, and simple explainability techniques. However, each of these choices also highlighted critical considerations for future studies aiming for deeper biological insight and predictive robustness. By refining preprocessing techniques, exploring advanced modeling frameworks, leveraging different generative and explainable AI methods, and more fully utilizing the longitudinal design of the experiment, future work can build upon the foundation laid here to uncover more nuanced, interpretable, and biologically meaningful patterns in bed rest and countermeasure research.

Conclusion

This thesis examined the effectiveness of machine learning in analyzing systemic physiological and genetic responses to simulated microgravity and nutritional interventions, utilizing data from the European Space Agency's Toulouse Cocktail HDT bed rest study. The aim was to assess whether a targeted antioxidant-rich nutritional supplement could mitigate the negative physiological consequences of mechanical unloading and whether biological recovery was evident at the end of the intervention. By integrating transcriptomic and core physiological data, we were able to assess complex biological interactions that span multiple body systems over time.

Through the application of supervised machine learning techniques – K-Nearest Neighbors, Random Forests, and Neural Networks – we classified subjects across different experimental timepoints and intervention conditions. We tested three biologically motivated hypotheses centered around the uniformity of intervention response, the distinction between baseline and altered physiological states during bed rest, and recovery to baseline at the end of the study.

Our findings indicate that the nutritional intervention did not have a measurably uniform effect across subjects, neither in gene expression profiles nor physiological features. These results, supported by multiple experiments and consistent across datasets, provide sufficient evidence for rejecting Hypothesis 1. The variation observed across individuals suggests that even well-formulated countermeasures may not yield uniform physiological effects, particularly when tested across small sample sizes with inherent biological variability.

Hypothesis 2, which posited a clear distinction between baseline (BDC1) and out-of-phase timepoints (HDT60, R1, and R3), was supported by high classification accuracies. This finding demonstrates that measurable systemic changes in physiology and gene expression occur during the bed rest period, reflecting the impact of microgravity analog conditions on the human body. These findings contribute to our understanding of how even short-term unloading environments can drive significant biological shifts.

Contrary to Hypothesis 3, we found that subjects had not universally returned to baseline physiological or genetic states by the end of the reambulation phase (R4). Despite a partial normalization in some features, classification models could still reliably distinguish R4 from

BDC1, indicating residual effects of the unloading period. These findings highlight the potential need for longer recovery periods or additional interventions post-bed rest or spaceflight.

From a biological perspective, these findings underscore the complexity and individuality of physiological adaptation to space analog conditions. While spaceflight countermeasures are often evaluated in isolation within single physiological systems, our results support the need for a more integrated, multi-system approach. The lack of systemic effects from the nutritional cocktail further raises important questions about the sufficiency of antioxidant-based interventions to mitigate widespread deconditioning.

Methodologically, this work highlights the potential of machine learning to uncover the presence or absence of meaningful patterns in high-dimensional biomedical data with limited sample sizes – a common limitation in space health research. The preprocessing decisions made in this thesis (such as robust scaling, targeted feature selection, and preservation of biological relevance) allowed models to operate effectively without sacrificing interpretability. The synthetic data experiments further reinforced our conclusions, illustrating that classifiability can depend more on the distinctiveness of class distributions than on the sample size.

This thesis lays the foundation for future work that could integrate additional omics layers, explore causal relationships, and implement explainability tools, such as SHAP values, for improved biological interpretation. As human spaceflight progresses toward longer missions and increased exposure to extreme environments, such integrative and data-driven approaches will be crucial for safeguarding astronaut health and optimizing countermeasure strategies.

Appendix

Appendix I: Cocktail Composition

This cocktail of antioxidant and anti-inflammatory agents is referred to as XXS-2A-BR2 and was designed by Spiral Company (Dijon, France).

A bioactive polyphenol compound (flavonols: 323.4 mg, phenylpropanoids: 45.6 mg, oligostilbenes: 78.0 mg, hydroxycinnamic acids: 50.4 mg, flavanols: 135.6 mg, flavanones: 108.0 mg) (XXS-2A-BR2; Spiral Company, Dijon, France) was daily administered, divided into 6 pills per day (equivalent to a daily dose of 741 mg), with 2 pills at breakfast, 2 at lunch and 2 at dinner, consisting of 168 mg of vitamin E with 80 µg of selenium (Solgar, Marne-La-Vallée, France) and 2.1 g of omega-3 fatty acids (Omacor®, Pierre Fabre, Toulouse, France).

Appendix II: Classification into phases

This appendix presents the comprehensive classification results for all five day-pair combinations across three different gene transformation datasets. For each transformation, multiple machine learning models were evaluated under varying parameter settings. The performance of these models was assessed using five standard evaluation metrics: accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). The tables report the mean and standard deviation of these metrics across all classification runs, providing a quantitative comparison of model performance. Additionally, corresponding heatmaps are included to visually highlight metric trends and support intuitive interpretation. These visualizations serve to complement the numerical results by emphasizing patterns in performance across datasets, models, and classification tasks.

1. **BDC 1 vs BDC 2**

   a. Log2 dataset

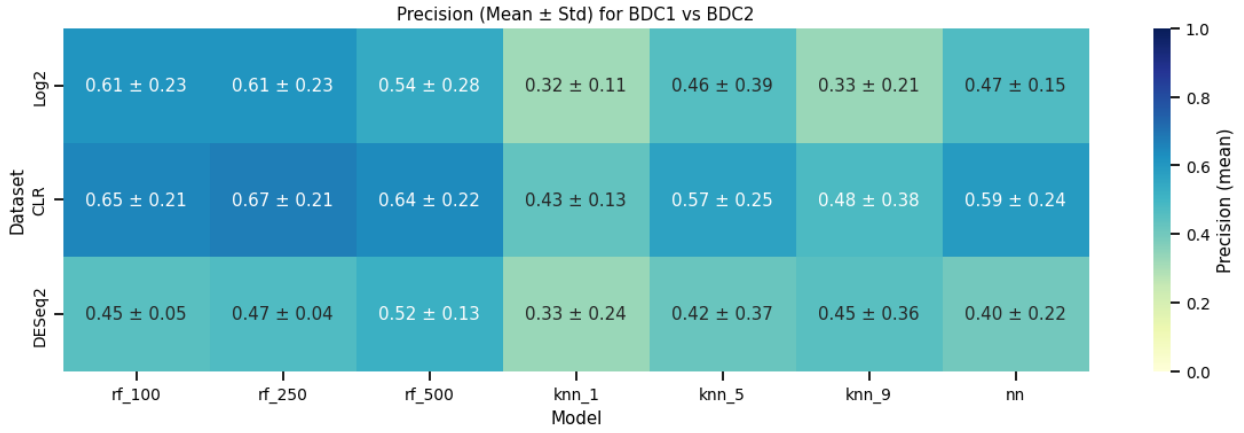
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.52	0.10	0.61	0.23	0.55	0.21	0.52	0.10	0.05	0.29
rf	250	0.52	0.10	0.61	0.23	0.50	0.25	0.49	0.13	0.05	0.29
rf	500	0.45	0.14	0.54	0.28	0.45	0.21	0.44	0.12	-0.10	0.35
knn	1	0.32	0.14	0.32	0.11	0.25	0.00	0.28	0.04	-0.36	0.29
knn	5	0.47	0.16	0.46	0.39	0.30	0.21	0.33	0.22	-0.02	0.36
knn	9	0.42	0.11	0.33	0.21	0.25	0.18	0.27	0.17	-0.15	0.22
nn		0.45	0.14	0.47	0.15	0.55	0.33	0.47	0.17	-0.12	0.31

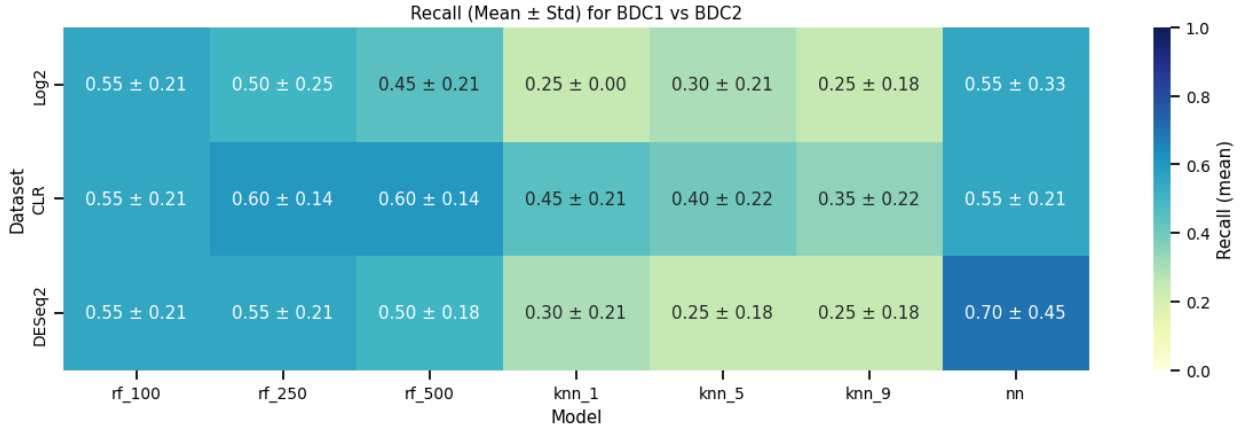
   b. CLR dataset

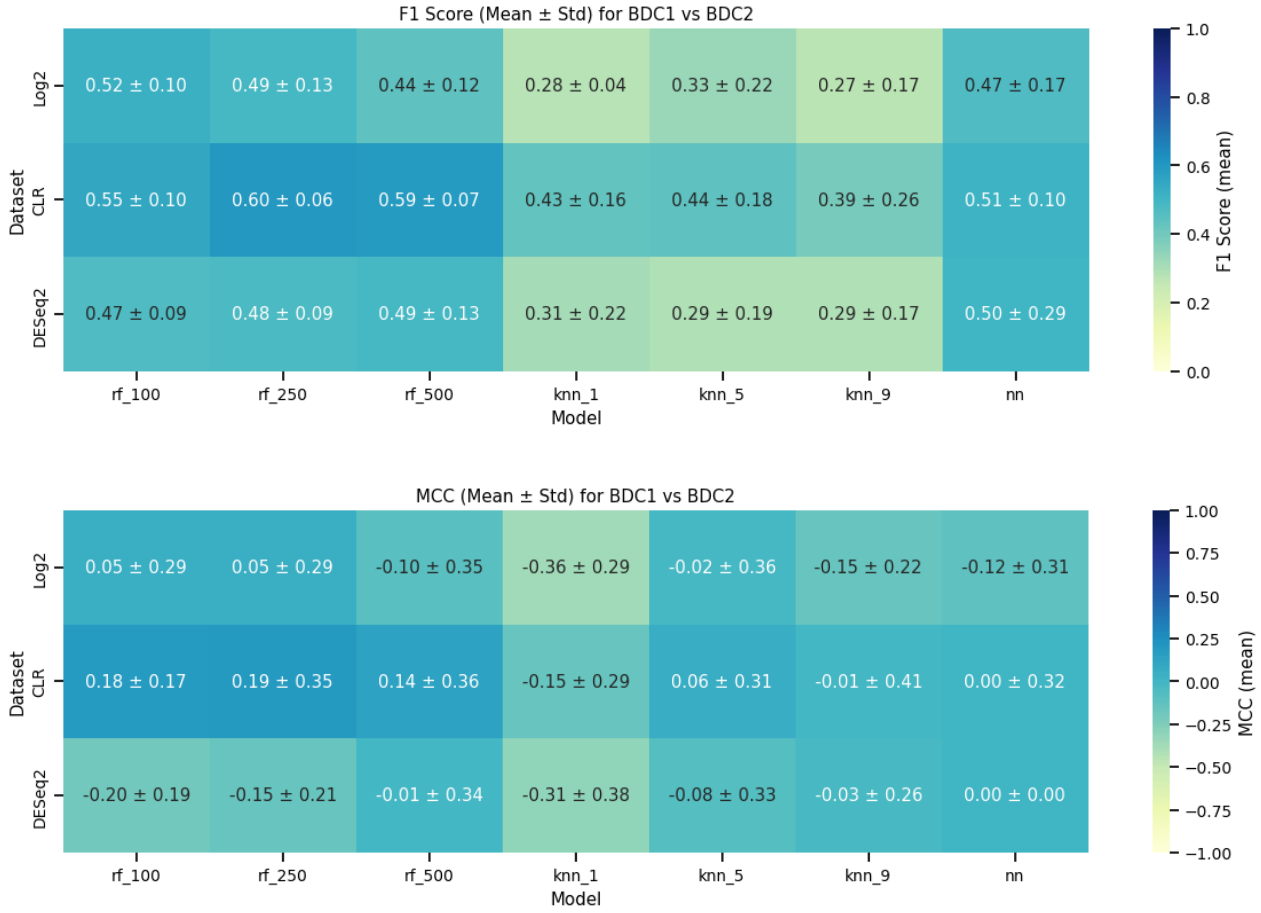
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.58	0.07	0.65	0.21	0.55	0.21	0.55	0.10	0.18	0.17
rf	250	0.60	0.14	0.67	0.21	0.60	0.14	0.60	0.06	0.19	0.35
rf	500	0.58	0.14	0.64	0.22	0.60	0.14	0.59	0.07	0.14	0.36
knn	1	0.43	0.14	0.43	0.13	0.45	0.21	0.43	0.16	-0.15	0.29
knn	5	0.53	0.14	0.57	0.25	0.40	0.22	0.44	0.18	0.06	0.31
knn	9	0.50	0.18	0.48	0.38	0.35	0.22	0.39	0.26	-0.01	0.41
nn		0.50	0.13	0.59	0.24	0.55	0.21	0.51	0.10	0.00	0.32

c. DESeq2 dataset

		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.43	0.07	0.45	0.05	0.55	0.21	0.47	0.09	-0.20	0.19
rf	250	0.45	0.07	0.47	0.04	0.55	0.21	0.48	0.09	-0.15	0.21
rf	500	0.50	0.15	0.52	0.13	0.50	0.18	0.49	0.13	-0.01	0.34
knn	1	0.35	0.19	0.33	0.24	0.30	0.21	0.31	0.22	-0.31	0.38
knn	5	0.45	0.14	0.42	0.37	0.25	0.18	0.29	0.19	-0.08	0.33
knn	9	0.48	0.10	0.45	0.36	0.25	0.18	0.29	0.17	-0.03	0.26
nn		0.50	0.00	0.40	0.22	0.70	0.45	0.50	0.29	0.00	0.00







Mean and standard deviation of classification metrics (Accuracy, Precision, Recall, F1 Score, and MCC) across different models (Random Forest, KNN, Neural Network) and data representations (Log2, CLR, DESeq2, and non-gene features) for distinguishing between BDC1 and BDC2 phases. Across all models and datasets, performance metrics remain close to chance level (e.g., ~0.5 accuracy), indicating that BDC1 and BDC2 are not classifiable supporting the hypothesis that no meaningful biological or physiological difference exists between these two pre-intervention time points.

2. BDC 1 vs HDT 60

a. Log2 dataset

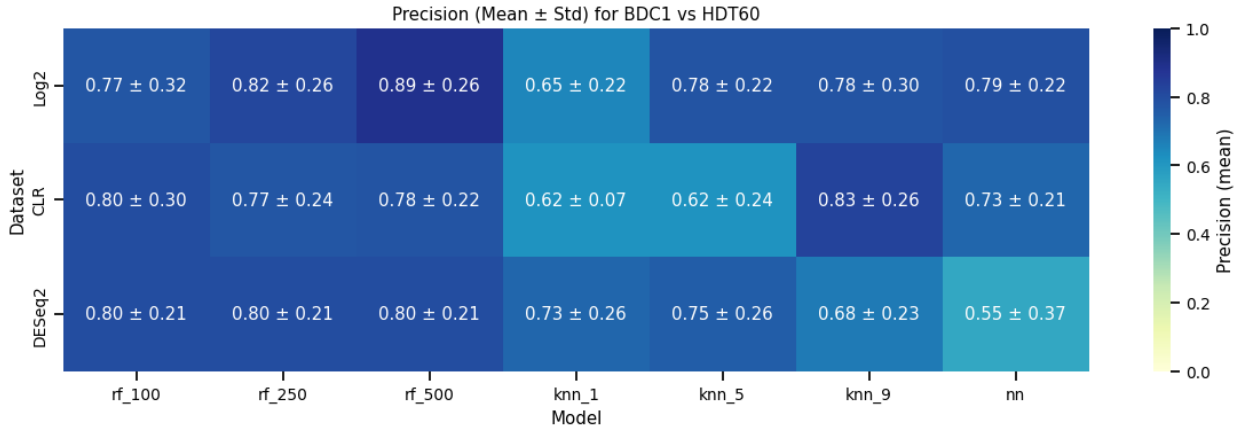
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.68	0.29	0.77	0.32	0.70	0.11	0.71	0.21	0.35	0.62
rf	250	0.73	0.24	0.82	0.26	0.70	0.21	0.73	0.20	0.45	0.54
rf	500	0.73	0.21	0.89	0.26	0.65	0.14	0.72	0.14	0.47	0.48
knn	1	0.60	0.14	0.65	0.22	0.60	0.14	0.60	0.11	0.22	0.30
knn	5	0.65	0.10	0.78	0.22	0.55	0.21	0.60	0.13	0.34	0.23
knn	9	0.60	0.16	0.78	0.30	0.40	0.14	0.50	0.16	0.25	0.37
nn		0.78	0.21	0.79	0.22	0.85	0.22	0.80	0.17	0.56	0.42

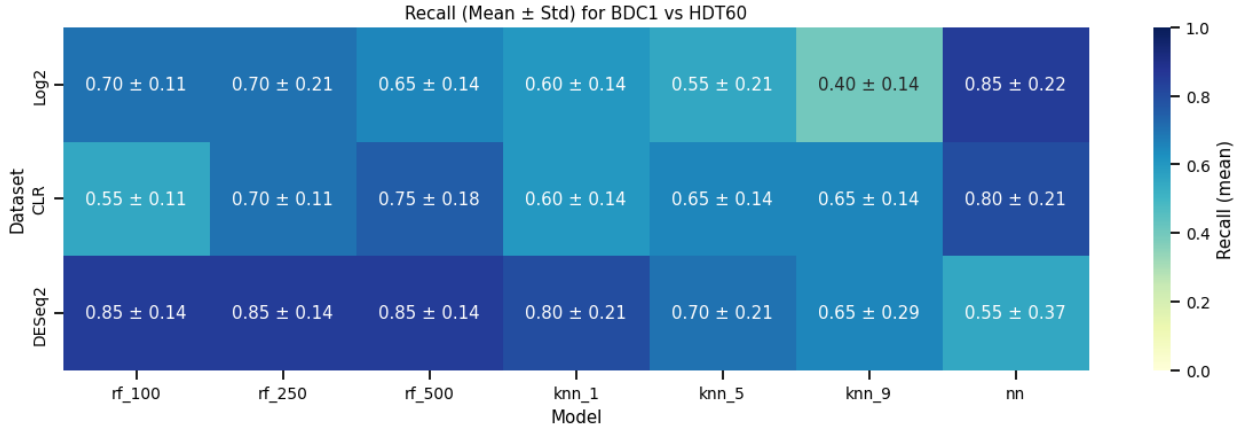
b. CLR dataset

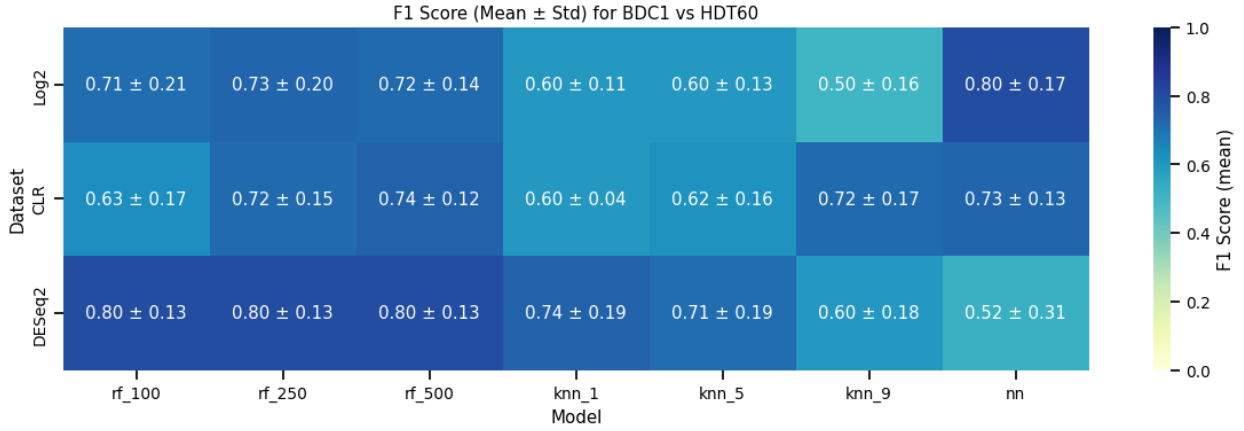
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.65	0.24	0.80	0.30	0.55	0.11	0.63	0.17	0.32	0.54
rf	250	0.70	0.21	0.77	0.24	0.70	0.11	0.72	0.15	0.39	0.48
rf	500	0.73	0.16	0.78	0.22	0.75	0.18	0.74	0.12	0.46	0.34
knn	1	0.60	0.06	0.62	0.07	0.60	0.14	0.60	0.04	0.21	0.12
knn	5	0.58	0.21	0.62	0.24	0.65	0.14	0.62	0.16	0.13	0.46
knn	9	0.73	0.21	0.83	0.26	0.65	0.14	0.72	0.17	0.47	0.43
nn		0.70	0.19	0.73	0.21	0.80	0.21	0.73	0.13	0.41	0.45

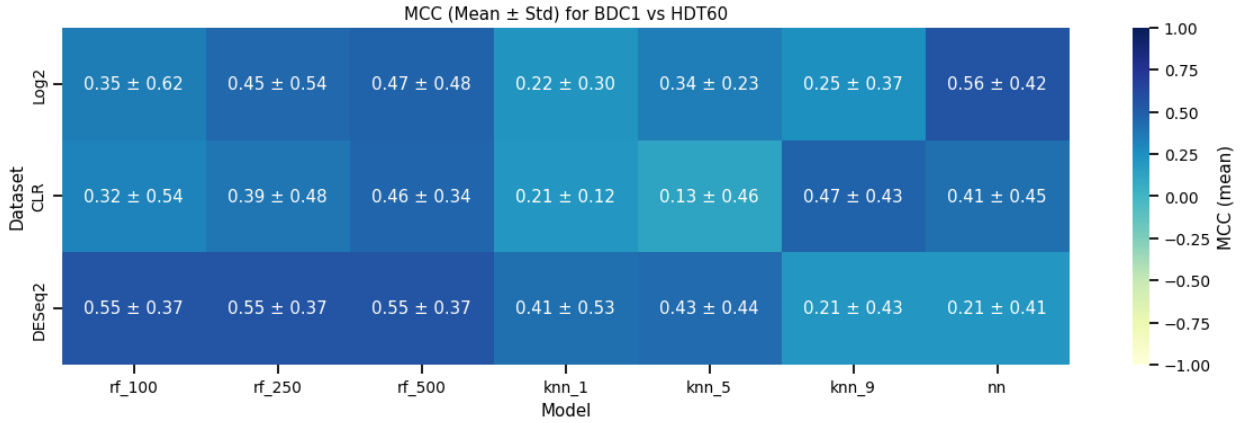
c. DESeq2 dataset

		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.78	0.19	0.80	0.21	0.85	0.14	0.80	0.13	0.55	0.37
rf	250	0.78	0.19	0.80	0.21	0.85	0.14	0.80	0.13	0.55	0.37
rf	500	0.78	0.19	0.80	0.21	0.85	0.14	0.80	0.13	0.55	0.37
knn	1	0.70	0.24	0.73	0.26	0.80	0.21	0.74	0.19	0.41	0.53
knn	5	0.70	0.21	0.75	0.26	0.70	0.21	0.71	0.19	0.43	0.44
knn	9	0.60	0.19	0.68	0.23	0.65	0.29	0.60	0.18	0.21	0.43
nn		0.60	0.16	0.55	0.37	0.55	0.37	0.52	0.31	0.21	0.41


Precision (Mean ± Std) for BDC1 vs HDT60
Dataset
Log2
CLR
DESeq2
0.77 ± 0.32
0.82 ± 0.26
0.89 ± 0.26
0.65 ± 0.22
0.78 ± 0.22
0.78 ± 0.30
0.79 ± 0.22
0.80 ± 0.30
0.77 ± 0.24
0.78 ± 0.22
0.62 ± 0.07
0.62 ± 0.24
0.83 ± 0.26
0.73 ± 0.21
0.80 ± 0.21
0.80 ± 0.21
0.80 ± 0.21
0.73 ± 0.26
0.75 ± 0.26
0.68 ± 0.23
0.55 ± 0.37
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
Precision (mean)
1.0
0.8
0.6
0.4
0.2
0.0


Recall (Mean ± Std) for BDC1 vs HDT60
Dataset
Log2
CLR
DESeq2
0.70 ± 0.11
0.70 ± 0.21
0.65 ± 0.14
0.60 ± 0.14
0.55 ± 0.21
0.40 ± 0.14
0.85 ± 0.22
0.55 ± 0.11
0.70 ± 0.11
0.75 ± 0.18
0.60 ± 0.14
0.65 ± 0.14
0.65 ± 0.14
0.80 ± 0.21
0.85 ± 0.14
0.85 ± 0.14
0.85 ± 0.14
0.80 ± 0.21
0.70 ± 0.21
0.65 ± 0.29
0.55 ± 0.37
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
Recall (mean)
1.0
0.8
0.6
0.4
0.2
0.0


F1 Score (Mean ± Std) for BDC1 vs HDT60
Dataset
Log2
CLR
DESeq2
0.71 ± 0.21
0.73 ± 0.20
0.72 ± 0.14
0.60 ± 0.11
0.60 ± 0.13
0.50 ± 0.16
0.80 ± 0.17
0.63 ± 0.17
0.72 ± 0.15
0.74 ± 0.12
0.60 ± 0.04
0.62 ± 0.16
0.72 ± 0.17
0.73 ± 0.13
0.80 ± 0.13
0.80 ± 0.13
0.80 ± 0.13
0.74 ± 0.19
0.71 ± 0.19
0.60 ± 0.18
0.52 ± 0.31
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
F1 Score (mean)
1.0
0.8
0.6
0.4
0.2
0.0



Mean and standard deviation of classification metrics (Accuracy, Precision, Recall, F1 Score, and MCC) across different models (Random Forest, KNN, Neural Network) and data representations (Log2, CLR, DESeq2, and non-gene features) for distinguishing between BDC1 and HDT 60 phases. Across all models and datasets, performance metrics remain above the chance level (e.g., ~0.5 accuracy), indicating that BDC1 and HDT 60 are classifiable supporting the hypothesis that no meaningful biological or physiological difference exists between these time points.

3. BDC 1 vs R1

a. Log2 dataset

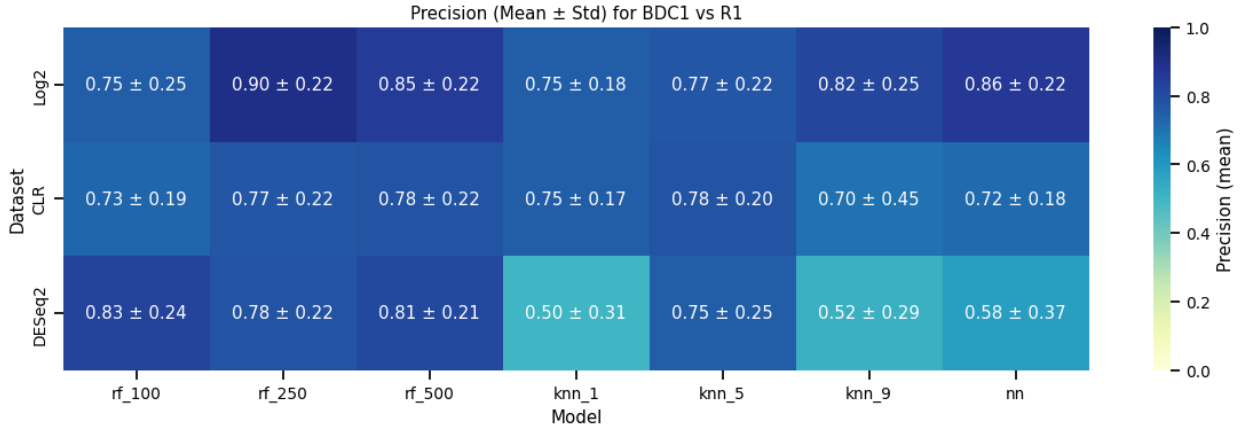
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.63	0.13	0.75	0.25	0.60	0.29	0.60	0.14	0.29	0.28
rf	250	0.75	0.20	0.90	0.22	0.70	0.33	0.72	0.23	0.55	0.38
rf	500	0.70	0.14	0.85	0.22	0.65	0.29	0.67	0.17	0.45	0.29
knn	1	0.75	0.15	0.75	0.18	0.75	0.35	0.72	0.23	0.54	0.32
knn	5	0.70	0.14	0.77	0.22	0.65	0.38	0.64	0.25	0.46	0.29
knn	9	0.65	0.14	0.82	0.25	0.45	0.27	0.53	0.22	0.36	0.28
nn		0.78	0.16	0.86	0.22	0.80	0.21	0.79	0.11	0.58	0.34

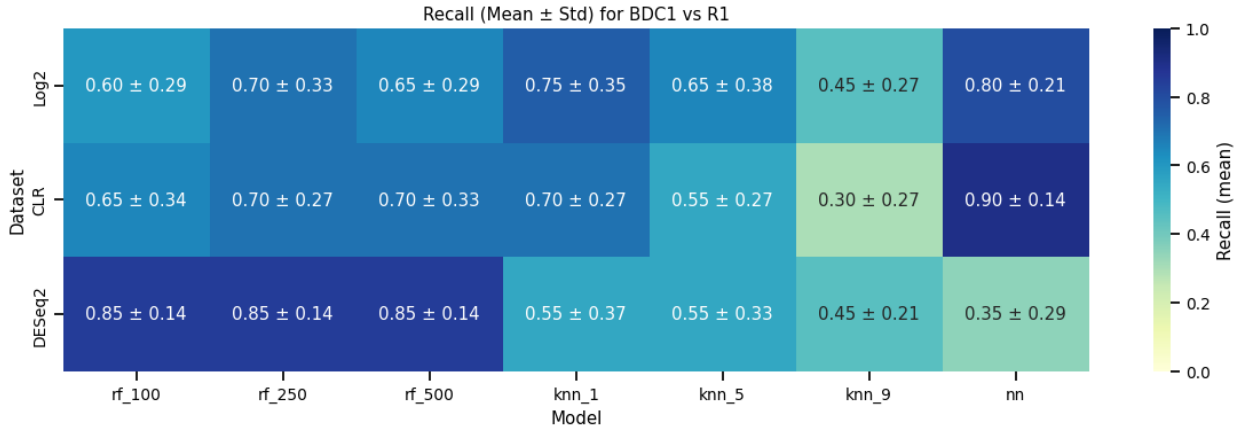
b. CLR dataset

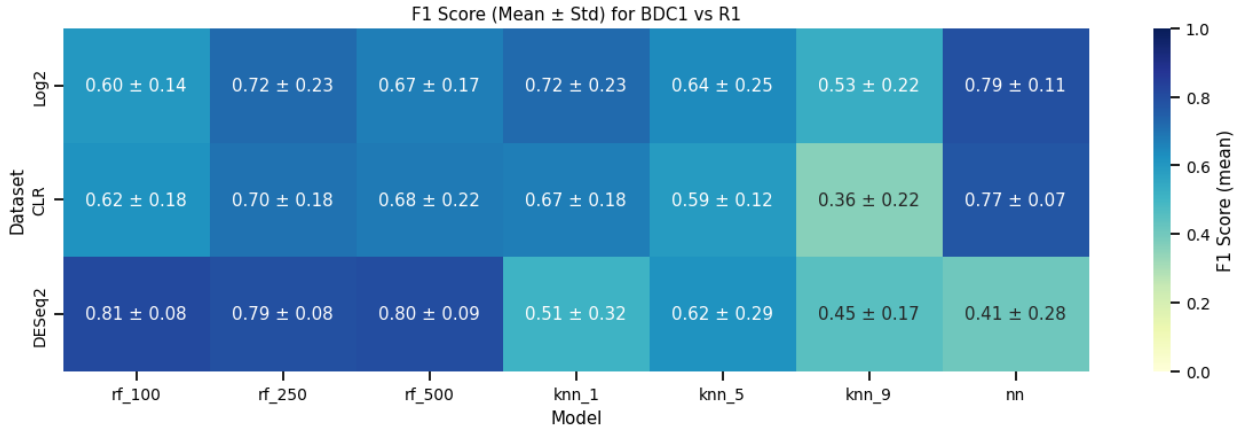
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.65	0.14	0.73	0.19	0.65	0.34	0.62	0.18	0.33	0.28
rf	250	0.70	0.19	0.77	0.22	0.70	0.27	0.70	0.18	0.42	0.38
rf	500	0.70	0.19	0.78	0.22	0.70	0.33	0.68	0.22	0.43	0.37
knn	1	0.70	0.11	0.75	0.17	0.70	0.27	0.67	0.18	0.43	0.22
knn	5	0.65	0.06	0.78	0.20	0.55	0.27	0.59	0.12	0.37	0.13
knn	9	0.55	0.11	0.70	0.45	0.30	0.27	0.36	0.22	0.15	0.34
nn		0.73	0.14	0.72	0.18	0.90	0.14	0.77	0.07	0.49	0.29

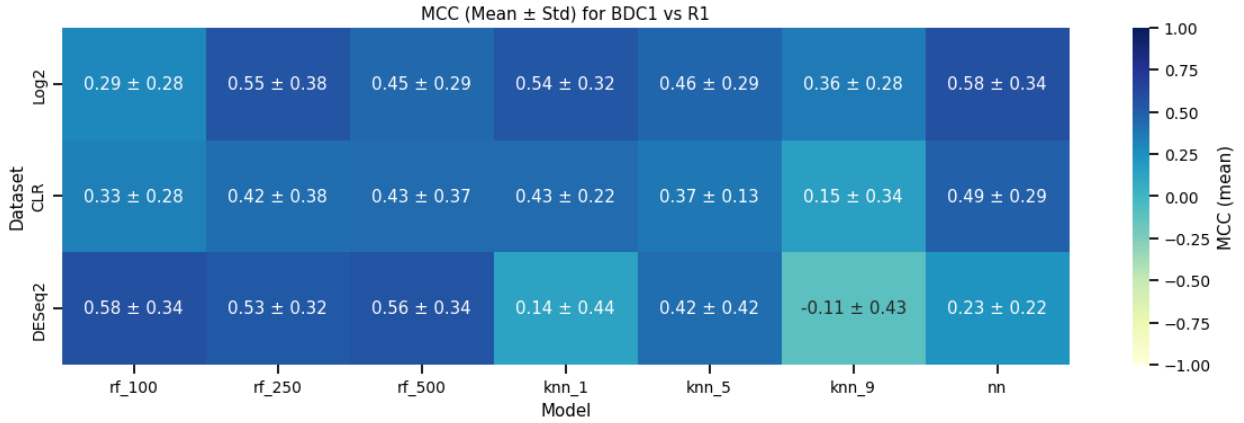
c. DESeq2 dataset

		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.78	0.16	0.83	0.24	0.85	0.14	0.81	0.08	0.58	0.34
rf	250	0.75	0.15	0.78	0.22	0.85	0.14	0.79	0.08	0.53	0.32
rf	500	0.78	0.16	0.81	0.21	0.85	0.14	0.80	0.09	0.56	0.34
knn	1	0.58	0.19	0.50	0.31	0.55	0.37	0.51	0.32	0.14	0.44
knn	5	0.70	0.21	0.75	0.25	0.55	0.33	0.62	0.29	0.42	0.42
knn	9	0.45	0.19	0.52	0.29	0.45	0.21	0.45	0.17	-0.11	0.43
nn		0.60	0.10	0.58	0.37	0.35	0.29	0.41	0.28	0.23	0.22









Mean and standard deviation of classification metrics (Accuracy, Precision, Recall, F1 Score, and MCC) across different models (Random Forest, KNN, Neural Network) and data representations (Log2, CLR, DESeq2, and non-gene features) for distinguishing between BDC1 and R1 phases. Across all models and datasets, performance metrics remain above the chance level (e.g., ~0.5

accuracy), indicating that BDC1 and R1 are classifiable supporting the hypothesis that no meaningful biological or physiological difference exists between these time points.

4. BDC 1 vs R3

a. Log2 dataset

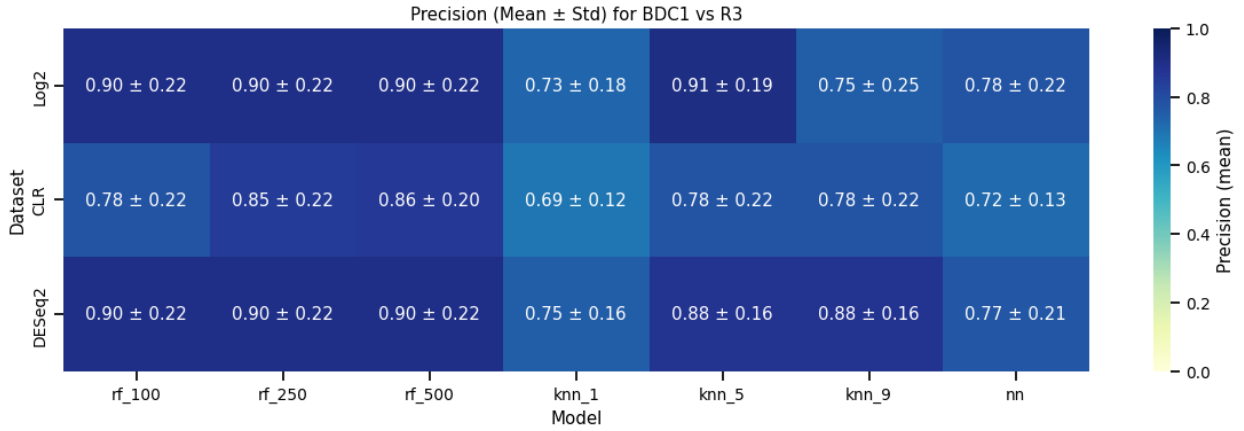
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.73	0.14	0.90	0.22	0.65	0.22	0.70	0.09	0.50	0.29
rf	250	0.73	0.14	0.90	0.22	0.65	0.22	0.70	0.09	0.50	0.29
rf	500	0.78	0.16	0.90	0.22	0.75	0.18	0.78	0.10	0.58	0.34
knn	1	0.70	0.14	0.73	0.18	0.70	0.11	0.71	0.12	0.41	0.29
knn	5	0.80	0.11	0.91	0.19	0.75	0.18	0.79	0.09	0.66	0.18
knn	9	0.63	0.13	0.75	0.25	0.60	0.29	0.60	0.14	0.29	0.28
nn		0.70	0.11	0.78	0.22	0.75	0.25	0.71	0.06	0.45	0.25

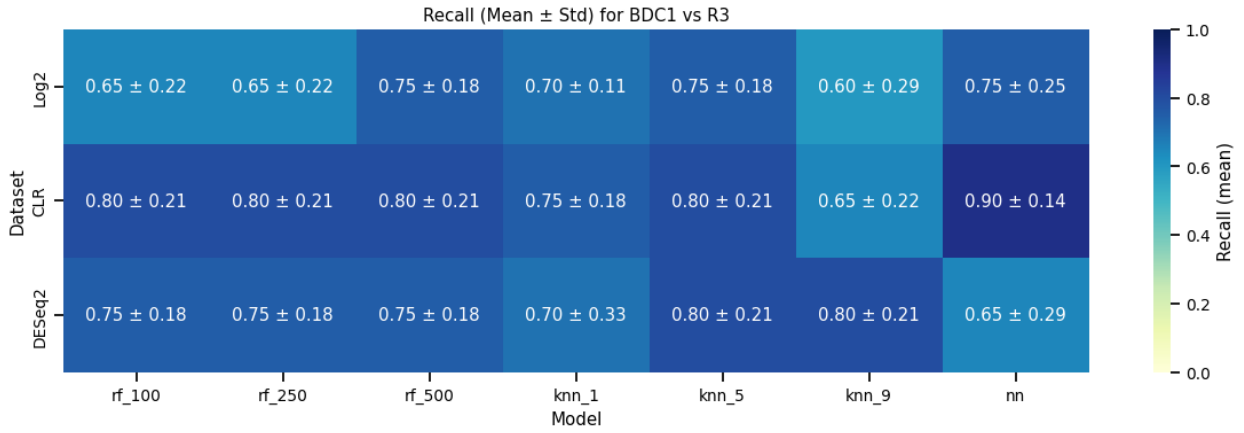
b. CLR dataset

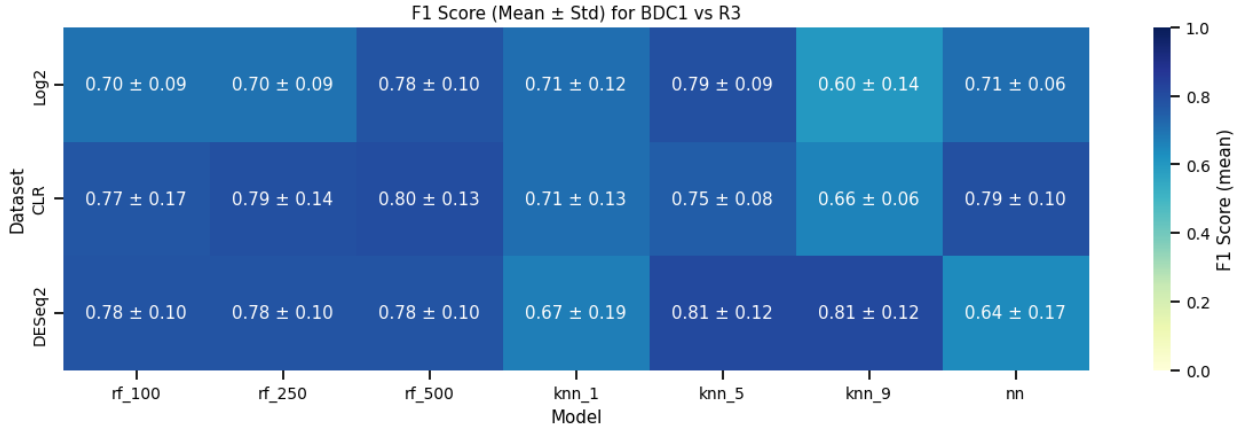
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.75	0.20	0.78	0.22	0.80	0.21	0.77	0.17	0.51	0.40
rf	250	0.78	0.19	0.85	0.22	0.80	0.21	0.79	0.14	0.57	0.37
rf	500	0.80	0.14	0.86	0.20	0.80	0.21	0.80	0.13	0.65	0.24
knn	1	0.70	0.14	0.69	0.12	0.75	0.18	0.71	0.13	0.41	0.29
knn	5	0.73	0.14	0.78	0.22	0.80	0.21	0.75	0.08	0.49	0.29
knn	9	0.68	0.11	0.78	0.22	0.65	0.22	0.66	0.06	0.38	0.25
nn		0.75	0.15	0.72	0.13	0.90	0.14	0.79	0.10	0.51	0.32

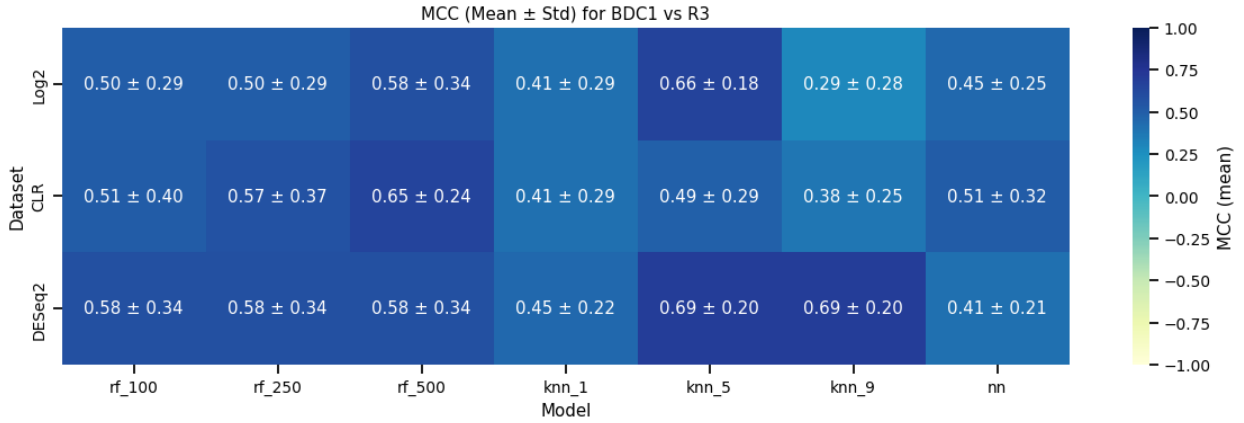
c. DESeq2 dataset

		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.78	0.16	0.90	0.22	0.75	0.18	0.78	0.10	0.58	0.34
rf	250	0.78	0.16	0.90	0.22	0.75	0.18	0.78	0.10	0.58	0.34
rf	500	0.78	0.16	0.90	0.22	0.75	0.18	0.78	0.10	0.58	0.34
knn	1	0.70	0.11	0.75	0.16	0.70	0.33	0.67	0.19	0.45	0.22
knn	5	0.83	0.11	0.88	0.16	0.80	0.21	0.81	0.12	0.69	0.20
knn	9	0.83	0.11	0.88	0.16	0.80	0.21	0.81	0.12	0.69	0.20
nn		0.68	0.11	0.77	0.21	0.65	0.29	0.64	0.17	0.41	0.21


Precision (Mean ± Std) for BDC1 vs R3
Dataset
Log2
CLR
DESeq2
0.90 ± 0.22
0.90 ± 0.22
0.90 ± 0.22
0.73 ± 0.18
0.91 ± 0.19
0.75 ± 0.25
0.78 ± 0.22
0.78 ± 0.22
0.85 ± 0.22
0.86 ± 0.20
0.69 ± 0.12
0.78 ± 0.22
0.78 ± 0.22
0.72 ± 0.13
0.90 ± 0.22
0.90 ± 0.22
0.90 ± 0.22
0.75 ± 0.16
0.88 ± 0.16
0.88 ± 0.16
0.77 ± 0.21
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
Precision (mean)
1.0
0.8
0.6
0.4
0.2
0.0


Recall (Mean ± Std) for BDC1 vs R3
Dataset
Log2
CLR
DESeq2
0.65 ± 0.22
0.65 ± 0.22
0.75 ± 0.18
0.70 ± 0.11
0.75 ± 0.18
0.60 ± 0.29
0.75 ± 0.25
0.80 ± 0.21
0.80 ± 0.21
0.80 ± 0.21
0.75 ± 0.18
0.80 ± 0.21
0.65 ± 0.22
0.90 ± 0.14
0.75 ± 0.18
0.75 ± 0.18
0.75 ± 0.18
0.70 ± 0.33
0.80 ± 0.21
0.80 ± 0.21
0.65 ± 0.29
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
Recall (mean)
1.0
0.8
0.6
0.4
0.2
0.0


F1 Score (Mean ± Std) for BDC1 vs R3
Dataset
Log2
CLR
DESeq2
0.70 ± 0.09
0.70 ± 0.09
0.78 ± 0.10
0.71 ± 0.12
0.79 ± 0.09
0.60 ± 0.14
0.71 ± 0.06
0.77 ± 0.17
0.79 ± 0.14
0.80 ± 0.13
0.71 ± 0.13
0.75 ± 0.08
0.66 ± 0.06
0.79 ± 0.10
0.78 ± 0.10
0.78 ± 0.10
0.78 ± 0.10
0.67 ± 0.19
0.81 ± 0.12
0.81 ± 0.12
0.64 ± 0.17
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
F1 Score (mean)
1.0
0.8
0.6
0.4
0.2
0.0



Mean and standard deviation of classification metrics (Accuracy, Precision, Recall, F1 Score, and MCC) across different models (Random Forest, KNN, Neural Network) and data representations (Log2, CLR, DESeq2, and non-gene features) for distinguishing between BDC1 and R3 phases. Across all models and datasets, performance metrics remain above the chance level (e.g., ~0.5 accuracy), indicating that BDC1 and R3 are classifiable supporting the hypothesis that no meaningful biological or physiological difference exists between these time points.

5. BDC 1 vs R4

a. Log2 dataset

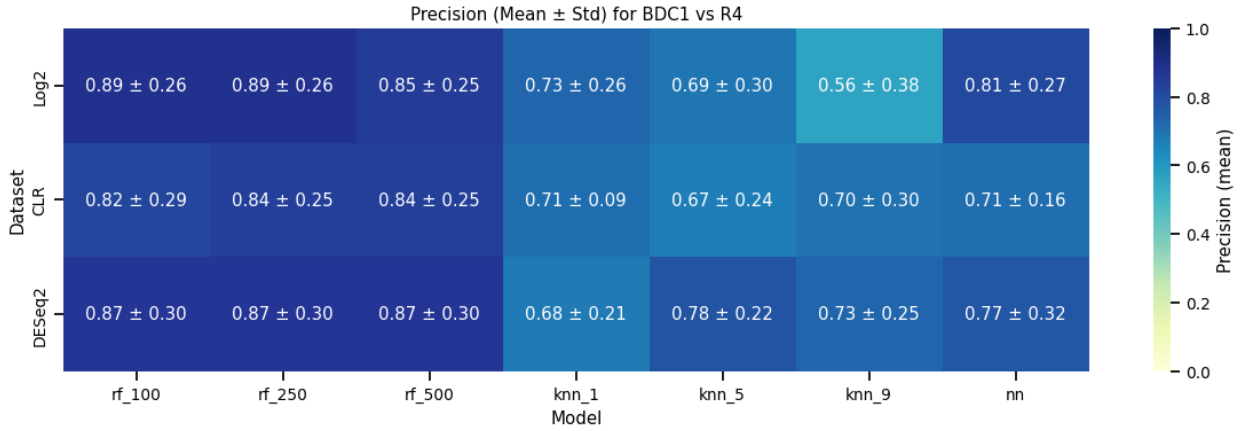
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.78	0.26	0.89	0.26	0.75	0.25	0.78	0.21	0.56	0.56
rf	250	0.75	0.27	0.89	0.26	0.70	0.33	0.72	0.27	0.52	0.57
rf	500	0.75	0.23	0.85	0.25	0.75	0.25	0.75	0.19	0.51	0.53
knn	1	0.63	0.18	0.73	0.26	0.55	0.21	0.59	0.18	0.28	0.37
knn	5	0.55	0.19	0.69	0.30	0.45	0.21	0.49	0.16	0.13	0.45
knn	9	0.60	0.16	0.56	0.38	0.45	0.27	0.49	0.29	0.22	0.35
nn		0.68	0.23	0.81	0.27	0.55	0.27	0.62	0.24	0.39	0.46

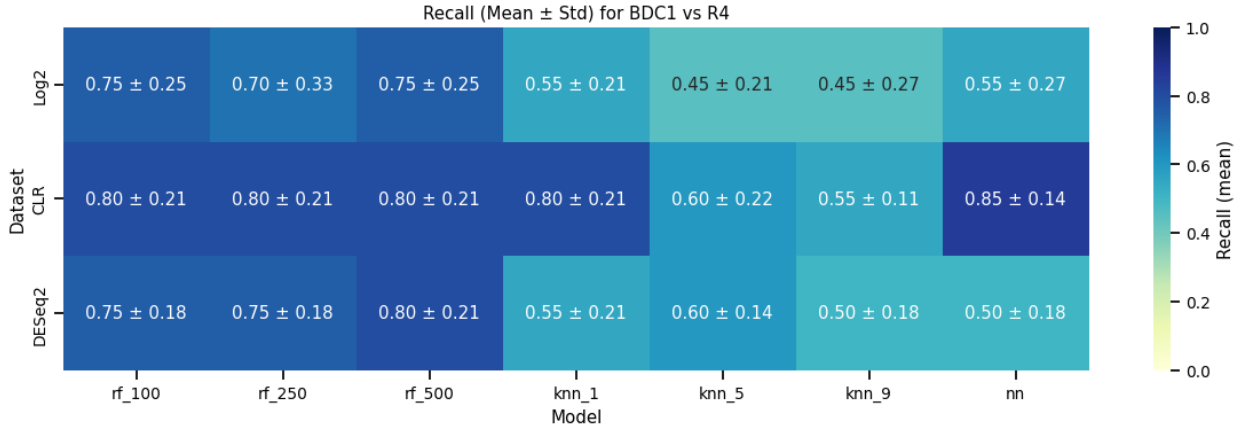
b. CLR dataset

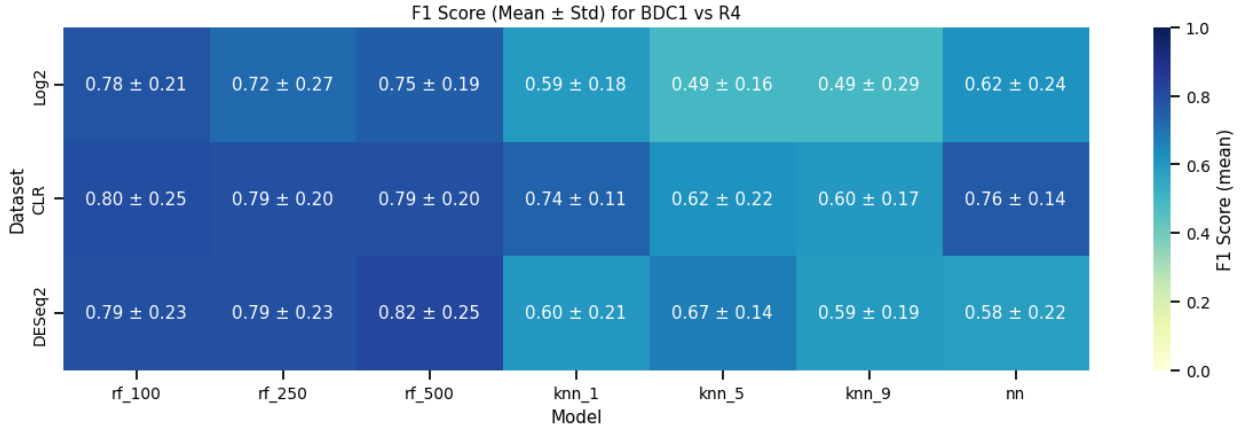
		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.78	0.31	0.82	0.29	0.80	0.21	0.80	0.25	0.54	0.66
rf	250	0.78	0.26	0.84	0.25	0.80	0.21	0.79	0.20	0.54	0.56
rf	500	0.78	0.26	0.84	0.25	0.80	0.21	0.79	0.20	0.54	0.56
knn	1	0.73	0.10	0.71	0.09	0.80	0.21	0.74	0.11	0.48	0.19
knn	5	0.63	0.27	0.67	0.24	0.60	0.22	0.62	0.22	0.24	0.56
knn	9	0.60	0.24	0.70	0.30	0.55	0.11	0.60	0.17	0.21	0.53
nn		0.73	0.21	0.71	0.16	0.85	0.14	0.76	0.14	0.43	0.47

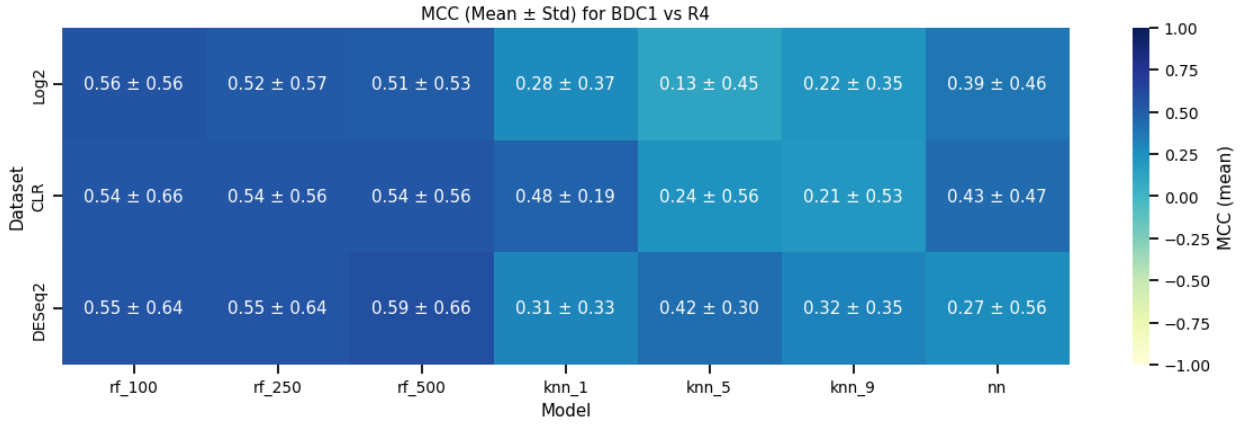
c. DESeq2 dataset

		Accuracy		Precision		Recall		F1 Score		MCC	
Model	**Param**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
rf	100	0.78	0.30	0.87	0.30	0.75	0.18	0.79	0.23	0.55	0.64
rf	250	0.78	0.30	0.87	0.30	0.75	0.18	0.79	0.23	0.55	0.64
rf	500	0.80	0.31	0.87	0.30	0.80	0.21	0.82	0.25	0.59	0.66
knn	1	0.65	0.16	0.68	0.21	0.55	0.21	0.60	0.21	0.31	0.33
knn	5	0.70	0.14	0.78	0.22	0.60	0.14	0.67	0.14	0.42	0.30
knn	9	0.65	0.16	0.73	0.25	0.50	0.18	0.59	0.19	0.32	0.35
nn		0.63	0.25	0.77	0.32	0.50	0.18	0.58	0.22	0.27	0.56


Precision (Mean ± Std) for BDC1 vs R4
Dataset
Log2
CLR
DESeq2
0.89 ± 0.26
0.89 ± 0.26
0.85 ± 0.25
0.73 ± 0.26
0.69 ± 0.30
0.56 ± 0.38
0.81 ± 0.27
0.82 ± 0.29
0.84 ± 0.25
0.84 ± 0.25
0.71 ± 0.09
0.67 ± 0.24
0.70 ± 0.30
0.71 ± 0.16
0.87 ± 0.30
0.87 ± 0.30
0.87 ± 0.30
0.68 ± 0.21
0.78 ± 0.22
0.73 ± 0.25
0.77 ± 0.32
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
Precision (mean)
1.0
0.8
0.6
0.4
0.2
0.0


Recall (Mean ± Std) for BDC1 vs R4
Dataset
Log2
CLR
DESeq2
0.75 ± 0.25
0.70 ± 0.33
0.75 ± 0.25
0.55 ± 0.21
0.45 ± 0.21
0.45 ± 0.27
0.55 ± 0.27
0.80 ± 0.21
0.80 ± 0.21
0.80 ± 0.21
0.80 ± 0.21
0.60 ± 0.22
0.55 ± 0.11
0.85 ± 0.14
0.75 ± 0.18
0.75 ± 0.18
0.80 ± 0.21
0.55 ± 0.21
0.60 ± 0.14
0.50 ± 0.18
0.50 ± 0.18
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
Recall (mean)
1.0
0.8
0.6
0.4
0.2
0.0


F1 Score (Mean ± Std) for BDC1 vs R4
Dataset
Log2
CLR
DESeq2
0.78 ± 0.21
0.72 ± 0.27
0.75 ± 0.19
0.59 ± 0.18
0.49 ± 0.16
0.49 ± 0.29
0.62 ± 0.24
0.80 ± 0.25
0.79 ± 0.20
0.79 ± 0.20
0.74 ± 0.11
0.62 ± 0.22
0.60 ± 0.17
0.76 ± 0.14
0.79 ± 0.23
0.79 ± 0.23
0.82 ± 0.25
0.60 ± 0.21
0.67 ± 0.14
0.59 ± 0.19
0.58 ± 0.22
rf_100
rf_250
rf_500
knn_1
knn_5
knn_9
nn
Model
F1 Score (mean)
1.0
0.8
0.6
0.4
0.2
0.0



Mean and standard deviation of classification metrics (Accuracy, Precision, Recall, F1 Score, and MCC) across different models (Random Forest, KNN, Neural Network) and data representations (Log2, CLR, DESeq2, and non-gene features) for distinguishing between BDC1 and R4 phases. Across all models and datasets, performance metrics remain above the chance level (e.g., ~0.5 accuracy), indicating that BDC1 and R4 are classifiable supporting the hypothesis that no meaningful biological or physiological difference exists between these time points.

The table below shows the intersection of genes that were ranked among top 20 in the feature importance scores for the day-pair classification.

BDC1 vs HDT60 - DESeq2 ∩ BDC1 vs HDT60 - Log2	ENSG00000156795, ENSG00000163002, ENSG00000181929	3
BDC1 vs R3 - CLR ∩ BDC1 vs R4 - CLR	ENSG00000179611, ENSG00000229180, ENSG00000232149	3
BDC1 vs R3 - CLR ∩ BDC1 vs R3 - DESeq2	ENSG00000109686, ENSG00000269570	2
BDC1 vs R3 - DESeq2 ∩ BDC1 vs R3 - Log2 ∩ BDC1 vs R4 - Log2	ENSG00000186714, ENSG00000249532	2
BDC1 vs R3 - DESeq2 ∩ BDC1 vs R4 - DESeq2	ENSG00000164889, ENSG00000178950	2
BDC1 vs R3 - Log2 ∩ BDC1 vs R4 - Log2	ENSG00000176973, ENSG00000257342	2
BDC1 vs HDT60 - CLR ∩ BDC1 vs R3 - DESeq2	ENSG00000106993, ENSG00000171155	2
BDC1 vs R3 - CLR ∩ BDC1 vs R4 - CLR ∩ BDC1 vs R4 - DESeq2	ENSG00000112309, ENSG00000271980	2
BDC1 vs R1 - DESeq2 ∩ BDC1 vs R4 - DESeq2	ENSG00000105556	1
BDC1 vs HDT60 - DESeq2 ∩ BDC1 vs HDT60 - Log2 ∩ BDC1 vs R4 - DESeq2	ENSG00000110090	1
BDC1 vs HDT60 - DESeq2 ∩ BDC1 vs R1 - DESeq2	ENSG00000110955	1
BDC1 vs R3 - CLR ∩ BDC1 vs R3 - DESeq2 ∩ BDC1 vs R3 - Log2 ∩ BDC1 vs R4 - DESeq2 ∩ BDC1 vs R4 - Log2	ENSG00000245970	1
BDC1 vs R3 - CLR ∩ BDC1 vs R3 - DESeq2 ∩ BDC1 vs R3 - Log2 ∩ BDC1 vs R4 - DESeq2	ENSG00000262160	1

BDC1 vs R3 - CLR ∩ BDC1 vs R3 - DESeq2 ∩ BDC1 vs R4 - CLR ∩ BDC1 vs R4 - DESeq2 ∩ BDC1 vs R4 - Log2	ENSG00000212802	1
BDC1 vs HDT60 - DESeq2 ∩ BDC1 vs R1 - Log2	ENSG00000158488	1
BDC1 vs HDT60 - Log2 ∩ BDC1 vs R4 - Log2	ENSG00000267576	1
BDC1 vs HDT60 - CLR ∩ BDC1 vs R1 - CLR	ENSG00000157445	1
BDC1 vs R3 - Log2 ∩ BDC1 vs R4 - DESeq2 ∩ BDC1 vs R4 - Log2	ENSG00000225798	1
BDC1 vs HDT60 - Log2 ∩ BDC1 vs R3 - CLR ∩ BDC1 vs R3 - Log2	ENSG00000221792	1
BDC1 vs R1 - DESeq2 ∩ BDC1 vs R3 - DESeq2	ENSG00000103197	1
BDC1 vs R1 - DESeq2 ∩ BDC1 vs R1 - Log2	ENSG00000175262	1
BDC1 vs HDT60 - CLR ∩ BDC1 vs R3 - CLR	ENSG00000129055	1
BDC1 vs R3 - DESeq2 ∩ BDC1 vs R3 - Log2	ENSG00000260934	1
BDC1 vs HDT60 - CLR ∩ BDC1 vs R1 - CLR ∩ BDC1 vs R1 - Log2	ENSG00000038219	1
BDC1 vs HDT60 - Log2 ∩ BDC1 vs R1 - CLR	ENSG00000270137	1
BDC1 vs R3 - CLR ∩ BDC1 vs R3 - Log2 ∩ BDC1 vs R4 - DESeq2	ENSG00000269246	1
BDC1 vs R4 - CLR ∩ BDC1 vs R4 - DESeq2	ENSG00000136709	1

Appendix III: Synthetic Data Results

1. Synthetic Gene Data – Sample Size 20

rf 50

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	50	1.00E-01	1	1	1	1	1
1	rf	50	1.00E-05	1	1	1	1	1
2	rf	50	1.00E-10	0	0	0	0	-1
3	rf	50	1.00E-20	0.5	0	0	0	0
4	rf	50	1.00E-30	1	1	1	1	1
5	rf	50	1.00E-40	0.25	0	0	0	-0.57735
6	rf	50	1.00E-50	0.25	0	0	0	-0.57735
7	rf	50	1.00E-60	0.75	0.666667	1	0.8	0.57735

rf 100

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	100	1.00E-01	1	1	1	1	1
1	rf	100	1.00E-05	1	1	1	1	1
2	rf	100	1.00E-10	0.25	0	0	0	-0.57735
3	rf	100	1.00E-20	0.75	1	0.5	0.666667	0.57735
4	rf	100	1.00E-30	0.25	0	0	0	-0.57735
5	rf	100	1.00E-40	0.25	0.333333	0.5	0.4	-0.57735
6	rf	100	1.00E-50	0.5	0.5	0.5	0.5	0
7	rf	100	1.00E-60	0.5	0.5	0.5	0.5	0

rf 150

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	150	1.00E-01	1	1	1	1	1
1	rf	150	1.00E-05	1	1	1	1	1
2	rf	150	1.00E-10	0.75	0.666667	1	0.8	0.57735
3	rf	150	1.00E-20	0.75	1	0.5	0.666667	0.57735
4	rf	150	1.00E-30	0.5	0.5	0.5	0.5	0
5	rf	150	1.00E-40	0.25	0	0	0	-0.57735
6	rf	150	1.00E-50	0.5	0.5	0.5	0.5	0
7	rf	150	1.00E-60	0.5	0	0	0	0

knn 1

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	1	1.00E-01	1	1	1	1	1
1	knn	1	1.00E-05	1	1	1	1	1
2	knn	1	1.00E-10	0.75	0.666667	1	0.8	0.57735
3	knn	1	1.00E-20	0.5	0.5	1	0.666667	0
4	knn	1	1.00E-30	0.5	0.5	0.5	0.5	0
5	knn	1	1.00E-40	0.25	0.333333	0.5	0.4	-0.57735
6	knn	1	1.00E-50	0.25	0.333333	0.5	0.4	-0.57735
7	knn	1	1.00E-60	0.75	0.666667	1	0.8	0.57735

Modek = knn 5

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	5	1.00E-01	1	1	1	1	1
1	knn	5	1.00E-05	1	1	1	1	1
2	knn	5	1.00E-10	0.5	0.5	1	0.666667	0
3	knn	5	1.00E-20	0.5	0.5	1	0.666667	0
4	knn	5	1.00E-30	0.75	0.666667	1	0.8	0.57735
5	knn	5	1.00E-40	0.5	0	0	0	0
6	knn	5	1.00E-50	0.25	0.333333	0.5	0.4	-0.57735
7	knn	5	1.00E-60	0.5	0	0	0	0

Model knn = 9

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	9	1.00E-01	1	1	1	1	1
1	knn	9	1.00E-05	1	1	1	1	1
2	knn	9	1.00E-10	0.5	0.5	0.5	0.5	0
3	knn	9	1.00E-20	0	0	0	0	-1
4	knn	9	1.00E-30	1	1	1	1	1
5	knn	9	1.00E-40	0.25	0	0	0	-0.57735
6	knn	9	1.00E-50	0.5	0.5	0.5	0.5	0
7	knn	9	1.00E-60	0.5	0	0	0	0

Model = nn

	model	param	a	accuracy	precision	recall	f1	mcc
0	nn	None	1.00E-01	1	1	1	1	1
1	nn	None	1.00E-05	1	1	1	1	1
2	nn	None	1.00E-10	0.5	0	0	0	0
3	nn	None	1.00E-20	0.5	0	0	0	0
4	nn	None	1.00E-30	0.5	0.5	1	0.666667	0
5	nn	None	1.00E-40	0.5	0.5	1	0.666667	0
6	nn	None	1.00E-50	0.5	0.5	1	0.666667	0
7	nn	None	1.00E-60	0.5	0	0	0	0

2. Synthetic Gene Data – Sample Size 200

rf 50

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	50	1.00E-01	1	1	1	1	1
1	rf	50	1.00E-05	1	1	1	1	1
2	rf	50	1.00E-10	0.8	0.833333	0.75	0.789474	0.603023
3	rf	50	1.00E-20	0.6	0.6	0.6	0.6	0.2
4	rf	50	1.00E-30	0.425	0.428571	0.45	0.439024	-0.150188
5	rf	50	1.00E-40	0.4	0.416667	0.5	0.454545	-0.204124
6	rf	50	1.00E-50	0.525	0.545455	0.3	0.387097	0.055989
7	rf	50	1.00E-60	0.525	0.521739	0.6	0.55814	0.050572

rf 100

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	100	1.00E-01	1	1	1	1	1
1	rf	100	1.00E-05	1	1	1	1	1
2	rf	100	1.00E-10	0.625	0.666667	0.5	0.571429	0.258199
3	rf	100	1.00E-20	0.525	0.52381	0.55	0.536585	0.050063
4	rf	100	1.00E-30	0.475	0.473684	0.45	0.461538	-0.050063
5	rf	100	1.00E-40	0.55	0.535714	0.75	0.625	0.109109
6	rf	100	1.00E-50	0.475	0.47619	0.5	0.487805	-0.050063
7	rf	100	1.00E-60	0.5	0.5	0.6	0.545455	0

rf 150

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	150	1.00E-01	1	1	1	1	1
1	rf	150	1.00E-05	1	1	1	1	1
2	rf	150	1.00E-10	0.775	0.73913	0.85	0.790698	0.556294
3	rf	150	1.00E-20	0.4	0.416667	0.5	0.454545	-0.204124
4	rf	150	1.00E-30	0.35	0.375	0.45	0.409091	-0.306186
5	rf	150	1.00E-40	0.525	0.521739	0.6	0.55814	0.050572
6	rf	150	1.00E-50	0.5	0.5	0.5	0.5	0
7	rf	150	1.00E-60	0.425	0.421053	0.4	0.410256	-0.15018

knn 1

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	1	1.00E-01	1	1	1	1	1
1	knn	1	1.00E-05	1	1	1	1	1
2	knn	1	1.00E-10	0.575	0.588235	0.5	0.540541	0.151717
3	knn	1	1.00E-20	0.575	0.615385	0.4	0.484848	0.160128
4	knn	1	1.00E-30	0.625	0.608696	0.7	0.651163	0.252861
5	knn	1	1.00E-40	0.375	0.391304	0.45	0.418605	-0.252861
6	knn	1	1.00E-50	0.55	0.55	0.55	0.55	0.1
7	knn	1	1.00E-60	0.525	0.533333	0.4	0.457143	0.05164

knn 5

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	5	1.00E-01	1	1	1	1	1
1	knn	5	1.00E-05	1	1	1	1	1
2	knn	5	1.00E-10	0.375	0.368421	0.35	0.358974	-0.250313
3	knn	5	1.00E-20	0.525	0.52381	0.55	0.536585	0.050063
4	knn	5	1.00E-30	0.525	0.52381	0.55	0.536585	0.050063
5	knn	5	1.00E-40	0.55	0.5625	0.45	0.5	0.102062
6	knn	5	1.00E-50	0.575	0.565217	0.65	0.604651	0.151717
7	knn	5	1.00E-60	0.575	0.565217	0.65	0.604651	0.151717

knn 9

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	9	1.00E-01	1	1	1	1	1
1	knn	9	1.00E-05	1	1	1	1	1
2	knn	9	1.00E-10	0.625	0.666667	0.5	0.571429	0.258199
3	knn	9	1.00E-20	0.575	0.571429	0.6	0.585366	0.150188
4	knn	9	1.00E-30	0.425	0.421053	0.4	0.410256	-0.150188
5	knn	9	1.00E-40	0.5	0.5	0.5	0.5	0
6	knn	9	1.00E-50	0.525	0.526316	0.5	0.512821	0.050063
7	knn	9	1.00E-60	0.4	0.409091	0.45	0.428571	-0.201008

nn

	model	param	a	accuracy	precision	recall	f1	mcc
0	nn	None	1.00E-01	1	1	1	1	1
1	nn	None	1.00E-05	1	1	1	1	1
2	nn	None	1.00E-10	0.5	0	0	0	0
3	nn	None	1.00E-20	0.5	0.5	1	0.666667	0
4	nn	None	1.00E-30	0.5	0.5	1	0.666667	0
5	nn	None	1.00E-40	0.5	0.5	1	0.666667	0
6	nn	None	1.00E-50	0.5	0.5	1	0.666667	0
7	nn	None	1.00E-60	0.5	0.5	1	0.666667	0

3. Synthetic Gene Data – Sample Size 2000

rf 50

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	50	1.00E-01	1	1	1	1	1
1	rf	50	1.00E-05	1	1	1	1	1
2	rf	50	1.00E-10	0.9875	0.98995	0.985	0.987469	0.975012
3	rf	50	1.00E-20	0.4925	0.492063	0.465	0.478149	-0.015023
4	rf	50	1.00E-30	0.495	0.494845	0.48	0.48731	-0.010005
5	rf	50	1.00E-40	0.4775	0.47644	0.455	0.465473	-0.045046
6	rf	50	1.00E-50	0.52	0.521277	0.49	0.505155	0.040072
7	rf	50	1.00E-60	0.5175	0.519126	0.475	0.496084	0.035127

rf 100

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	100	1.00E-01	1	1	1	1	1
1	rf	100	1.00E-05	1	1	1	1	1
2	rf	100	1.00E-10	0.97	0.979592	0.96	0.969697	0.940188
3	rf	100	1.00E-20	0.4925	0.492823	0.515	0.503667	-0.015015
4	rf	100	1.00E-30	0.47	0.468085	0.44	0.453608	-0.060108
5	rf	100	1.00E-40	0.4775	0.477157	0.47	0.473552	-0.045005
6	rf	100	1.00E-50	0.4475	0.444444	0.42	0.431877	-0.105159
7	rf	100	1.00E-60	0.57	0.573684	0.545	0.558974	0.140175

rf 150

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	150	1.00E-01	1	1	1	1	1
1	rf	150	1.00E-05	1	1	1	1	1
2	rf	150	1.00E-10	0.995	0.990099	1	0.995025	0.99005
3	rf	150	1.00E-20	0.4975	0.497696	0.54	0.517986	-0.005018
4	rf	150	1.00E-30	0.46	0.461165	0.475	0.46798	-0.080036
5	rf	150	1.00E-40	0.485	0.484211	0.46	0.471795	-0.030038
6	rf	150	1.00E-50	0.49	0.489247	0.455	0.471503	-0.020049
7	rf	150	1.00E-60	0.47	0.469697	0.465	0.467337	-0.060003

knn 1

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	1	1.00E-01	1	1	1	1	1
1	knn	1	1.00E-05	1	1	1	1	1
2	knn	1	1.00E-10	0.4925	0.492308	0.48	0.486076	-0.015005
3	knn	1	1.00E-20	0.52	0.520619	0.505	0.51269	0.040018
4	knn	1	1.00E-30	0.5275	0.526829	0.54	0.533333	0.055017
5	knn	1	1.00E-40	0.4875	0.486339	0.445	0.464752	-0.025091
6	knn	1	1.00E-50	0.52	0.521277	0.49	0.505155	0.040072
7	knn	1	1.00E-60	0.52	0.519608	0.53	0.524752	0.040008

knn 5

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	5	1.00E-01	1	1	1	1	1
1	knn	5	1.00E-05	1	1	1	1	1
2	knn	5	1.00E-10	0.4875	0.486772	0.46	0.473008	-0.025038
3	knn	5	1.00E-20	0.5375	0.536946	0.545	0.540943	0.075008
4	knn	5	1.00E-30	0.5225	0.525714	0.46	0.490667	0.045356
5	knn	5	1.00E-40	0.49	0.490291	0.505	0.497537	-0.020009
6	knn	5	1.00E-50	0.5125	0.512821	0.5	0.506329	0.025008
7	knn	5	1.00E-60	0.5075	0.507772	0.49	0.498728	0.015009

knn 9

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	9	1.00E-01	1	1	1	1	1
1	knn	9	1.00E-05	1	1	1	1	1
2	knn	9	1.00E-10	0.5125	0.512438	0.515	0.513716	0.025
3	knn	9	1.00E-20	0.46	0.461538	0.48	0.470588	-0.080064
4	knn	9	1.00E-30	0.5375	0.536946	0.545	0.540943	0.075008
5	knn	9	1.00E-40	0.495	0.495	0.495	0.495	-0.01
6	knn	9	1.00E-50	0.5025	0.502703	0.465	0.483117	0.005014
7	knn	9	1.00E-60	0.5375	0.535211	0.57	0.552058	0.075159

nn

	model	param	a	accuracy	precision	recall	f1	mcc
0	nn	None	1.00E-01	1	1	1	1	1
1	nn	None	1.00E-05	1	1	1	1	1
2	nn	None	1.00E-10	0.5	0.5	1	0.666667	0
3	nn	None	1.00E-20	0.5	0	0	0	0
4	nn	None	1.00E-30	0.5	0.5	1	0.666667	0
5	nn	None	1.00E-40	0.5	0	0	0	0
6	nn	None	1.00E-50	0.5	0	0	0	0
7	nn	None	1.00E-60	0.5	0.5	1	0.666667	0

4. Synthetic Gene Data – Sample Size 2000

knn 1

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	1	1.00E-01	1	1	1	1	1
1	knn	1	1.00E-05	1	1	1	1	1
2	knn	1	1.00E-10	0.50025	0.500252	0.497	0.498621	0.0005
3	knn	1	1.00E-20	0.508	0.50813	0.5	0.504032	0.016002
4	knn	1	1.00E-30	0.49	0.49004	0.492	0.491018	-0.02
5	knn	1	1.00E-40	0.50125	0.501252	0.5005	0.500876	0.0025
6	knn	1	1.00E-50	0.48825	0.488185	0.4855	0.486839	-0.0235
7	knn	1	1.00E-60	0.48975	0.489578	0.4815	0.485505	-0.020503

knn 5

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	5	1.00E-01	1	1	1	1	1
1	knn	5	1.00E-05	1	1	1	1	1
2	knn	5	1.00E-10	0.5065	0.506429	0.512	0.509199	0.013001
3	knn	5	1.00E-20	0.5085	0.508586	0.5035	0.50603	0.017001
4	knn	5	1.00E-30	0.4935	0.493487	0.4925	0.492993	-0.013
5	knn	5	1.00E-40	0.502	0.501959	0.5125	0.507175	0.004001
6	knn	5	1.00E-50	0.502	0.501953	0.514	0.507905	0.004001
7	knn	5	1.00E-60	0.4955	0.495431	0.488	0.491688	-0.009001

knn 9

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	9	1.00E-01	1	1	1	1	1
1	knn	9	1.00E-05	1	1	1	1	1
2	knn	9	1.00E-10	0.50875	0.508763	0.508	0.508381	0.0175
3	knn	9	1.00E-20	0.5155	0.515562	0.5135	0.514529	0.031
4	knn	9	1.00E-30	0.50075	0.500738	0.509	0.504835	0.0015
5	knn	9	1.00E-40	0.50275	0.502692	0.5135	0.508039	0.005501
6	knn	9	1.00E-50	0.4985	0.498537	0.511	0.504691	-0.003001
7	knn	9	1.00E-60	0.49175	0.491446	0.474	0.482566	-0.01651

nn

	model	param	a	accuracy	precision	recall	f1	mcc
0	nn	None	1.00E-01	1	1	1	1	1
1	nn	None	1.00E-05	1	1	1	1	1
2	nn	None	1.00E-10	0.5	0.5	1	0.666667	0
3	nn	None	1.00E-20	0.5	0	0	0	0
4	nn	None	1.00E-30	0.5	0	0	0	0
5	nn	None	1.00E-40	0.5	0.5	1	0.666667	0
6	nn	None	1.00E-50	0.5	0.5	1	0.666667	0
7	nn	None	1.00E-60	0.5	0.5	1	0.666667	0

rf 50

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	50	1.00E-01	1	1	1	1	1
1	rf	50	1.00E-05	1	1	1	1	1
2	rf	50	1.00E-10	0.9875	0.98995	0.985	0.987469	0.975012
3	rf	50	1.00E-20	0.4925	0.492063	0.465	0.478149	-0.015023
4	rf	50	1.00E-30	0.495	0.494845	0.48	0.48731	-0.010005
5	rf	50	1.00E-40	0.4775	0.47644	0.455	0.465473	-0.045046
6	rf	50	1.00E-50	0.52	0.521277	0.49	0.505155	0.040072
7	rf	50	1.00E-60	0.5175	0.519126	0.475	0.496084	0.035127

rf 100

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	100	1.00E-01	1	1	1	1	1
1	rf	100	1.00E-05	1	1	1	1	1
2	rf	100	1.00E-10	0.97	0.979592	0.96	0.969697	0.940188
3	rf	100	1.00E-20	0.4925	0.492823	0.515	0.503667	-0.015015
4	rf	100	1.00E-30	0.47	0.468085	0.44	0.453608	-0.060108
5	rf	100	1.00E-40	0.4775	0.477157	0.47	0.473552	-0.045005
6	rf	100	1.00E-50	0.4475	0.444444	0.42	0.431877	-0.105159
7	rf	100	1.00E-60	0.57	0.573684	0.545	0.558974	0.140175

rf 150

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	150	1.00E-01	1	1	1	1	1
1	rf	150	1.00E-05	1	1	1	1	1
2	rf	150	1.00E-10	0.995	0.990099	1	0.995025	0.99005
3	rf	150	1.00E-20	0.4975	0.497696	0.54	0.517986	-0.005018
4	rf	150	1.00E-30	0.46	0.461165	0.475	0.46798	-0.080036
5	rf	150	1.00E-40	0.485	0.484211	0.46	0.471795	-0.030038
6	rf	150	1.00E-50	0.49	0.489247	0.455	0.471503	-0.020049
7	rf	150	1.00E-60	0.47	0.469697	0.465	0.467337	-0.060003

5. Synthetic Feature data – sample size 20

rf 50

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	50	1.00E-01	1	1	1	1	1
1	rf	50	1.00E-02	1	1	1	1	1
2	rf	50	1.00E-03	0.25	0.333333	0.5	0.4	-0.57735
3	rf	50	1.00E-04	0.5	0.5	0.5	0.5	0
4	rf	50	1.00E-05	0.25	0.333333	0.5	0.4	-0.57735
5	rf	50	1.00E-06	0.75	1	0.5	0.666667	0.57735
6	rf	50	1.00E-07	0.25	0.333333	0.5	0.4	-0.57735
7	rf	50	1.00E-08	0.5	0.5	0.5	0.5	0

rf 100

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	100	1.00E-01	1	1	1	1	1
1	rf	100	1.00E-02	1	1	1	1	1
2	rf	100	1.00E-03	0.5	0.5	1	0.666667	0
3	rf	100	1.00E-04	0.75	1	0.5	0.666667	0.57735
4	rf	100	1.00E-05	1	1	1	1	1
5	rf	100	1.00E-06	0.75	1	0.5	0.666667	0.57735
6	rf	100	1.00E-07	0.75	1	0.5	0.666667	0.57735
7	rf	100	1.00E-08	0.25	0	0	0	-0.57735

rf 150

model	param	a	accuracy	precision	recall	f1	mcc	
0	rf	150	1.00E-01	1	1	1	1	1
1	rf	150	1.00E-02	1	1	1	1	1
2	rf	150	1.00E-03	0.75	0.666667	1	0.8	0.57735
3	rf	150	1.00E-04	0.5	0	0	0	0
4	rf	150	1.00E-05	0.5	0.5	1	0.666667	0
5	rf	150	1.00E-06	0.25	0.333333	0.5	0.4	-0.57735
6	rf	150	1.00E-07	0.75	0.666667	1	0.8	0.57735
7	rf	150	1.00E-08	0.5	0.5	0.5	0.5	0

knn 1

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	1	1.00E-01	1	1	1	1	1
1	knn	1	1.00E-02	1	1	1	1	1
2	knn	1	1.00E-03	0.25	0	0	0	-0.57735
3	knn	1	1.00E-04	0.75	0.666667	1	0.8	0.57735
4	knn	1	1.00E-05	0.5	0.5	1	0.666667	0
5	knn	1	1.00E-06	0.5	0	0	0	0
6	knn	1	1.00E-07	0.5	0	0	0	0
7	knn	1	1.00E-08	0.25	0	0	0	-0.57735

knn 5

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	5	1.00E-01	1	1	1	1	1
1	knn	5	1.00E-02	1	1	1	1	1
2	knn	5	1.00E-03	0.75	1	0.5	0.666667	0.57735
3	knn	5	1.00E-04	0.75	0.666667	1	0.8	0.57735
4	knn	5	1.00E-05	0.25	0	0	0	-0.57735
5	knn	5	1.00E-06	0.25	0	0	0	-0.57735
6	knn	5	1.00E-07	0.5	0.5	1	0.666667	0
7	knn	5	1.00E-08	0.25	0.333333	0.5	0.4	-0.57735

knn 9

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	9	1.00E-01	1	1	1	1	1
1	knn	9	1.00E-02	1	1	1	1	1
2	knn	9	1.00E-03	0.75	0.666667	1	0.8	0.57735
3	knn	9	1.00E-04	0.5	0.5	1	0.666667	0
4	knn	9	1.00E-05	0.25	0	0	0	-0.57735
5	knn	9	1.00E-06	0.5	0.5	1	0.666667	0
6	knn	9	1.00E-07	0.5	0.5	1	0.666667	0
7	knn	9	1.00E-08	0.25	0	0	0	-0.57735

nn

	model	param	a	accuracy	precision	recall	f1	mcc
0	nn	None	1.00E-01	1	1	1	1	1
1	nn	None	1.00E-02	1	1	1	1	1
2	nn	None	1.00E-03	0.5	0.5	0.5	0.5	0
3	nn	None	1.00E-04	0.75	0.666667	1	0.8	0.57735
4	nn	None	1.00E-05	0.75	1	0.5	0.666667	0.57735
5	nn	None	1.00E-06	0.25	0	0	0	-0.57735
6	nn	None	1.00E-07	0.75	0.666667	1	0.8	0.57735
7	nn	None	1.00E-08	0.5	0.5	0.5	0.5	0

6. Synthetic Feature data – sample size 200

rf 50

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	50	1.00E-01	1	1	1	1	1
1	rf	50	1.00E-02	1	1	1	1	1
2	rf	50	1.00E-03	0.6	0.590909	0.65	0.619048	0.201008
3	rf	50	1.00E-04	0.525	0.526316	0.5	0.512821	0.050063
4	rf	50	1.00E-05	0.4	0.375	0.3	0.333333	-0.204124
5	rf	50	1.00E-06	0.6	0.583333	0.7	0.636364	0.204124
6	rf	50	1.00E-07	0.55	0.571429	0.4	0.470588	0.104828
7	rf	50	1.00E-08	0.5	0.5	0.4	0.444444	0

rf 100

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	100	1.00E-01	1	1	1	1	1
1	rf	100	1.00E-02	1	1	1	1	1
2	rf	100	1.00E-03	0.65	0.714286	0.5	0.588235	0.314485
3	rf	100	1.00E-04	0.525	0.52381	0.55	0.536585	0.050063
4	rf	100	1.00E-05	0.45	0.45	0.45	0.45	-0.1
5	rf	100	1.00E-06	0.475	0.47619	0.5	0.487805	-0.050063
6	rf	100	1.00E-07	0.45	0.444444	0.4	0.421053	-0.100504
7	rf	100	1.00E-08	0.45	0.454545	0.5	0.47619	-0.100504

rf 150

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	150	1.00E-01	1	1	1	1	1
1	rf	150	1.00E-02	1	1	1	1	1
2	rf	150	1.00E-03	0.55	0.571429	0.4	0.470588	0.104828
3	rf	150	1.00E-04	0.475	0.470588	0.4	0.432432	-0.050572
4	rf	150	1.00E-05	0.475	0.482759	0.7	0.571429	-0.055989
5	rf	150	1.00E-06	0.4	0.4	0.4	0.4	-0.2
6	rf	150	1.00E-07	0.55	0.55	0.55	0.55	0.1
7	rf	150	1.00E-08	0.65	0.636364	0.7	0.666667	0.301511

knn 1

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	1	1.00E-01	1	1	1	1	1
1	knn	1	1.00E-02	1	1	1	1	1
2	knn	1	1.00E-03	0.65	0.636364	0.7	0.666667	0.301511
3	knn	1	1.00E-04	0.65	0.607143	0.85	0.708333	0.327327
4	knn	1	1.00E-05	0.3	0.1	0.05	0.066667	-0.46188
5	knn	1	1.00E-06	0.4	0.409091	0.45	0.428571	-0.201008
6	knn	1	1.00E-07	0.525	0.529412	0.45	0.486486	0.050572
7	knn	1	1.00E-08	0.525	0.526316	0.5	0.512821	0.050063

knn 5

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	5	1.00E-01	1	1	1	1	1
1	knn	5	1.00E-02	1	1	1	1	1
2	knn	5	1.00E-03	0.625	0.777778	0.35	0.482759	0.299342
3	knn	5	1.00E-04	0.5	0.5	0.4	0.444444	0
4	knn	5	1.00E-05	0.525	0.52381	0.55	0.536585	0.050063
5	knn	5	1.00E-06	0.5	0.5	0.6	0.545455	0
6	knn	5	1.00E-07	0.55	0.555556	0.5	0.526316	0.100504
7	knn	5	1.00E-08	0.525	0.533333	0.4	0.457143	0.05164

knn 9

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	9	1.00E-01	1	1	1	1	1
1	knn	9	1.00E-02	1	1	1	1	1
2	knn	9	1.00E-03	0.525	0.538462	0.35	0.424242	0.053376
3	knn	9	1.00E-04	0.425	0.428571	0.45	0.439024	-0.150188
4	knn	9	1.00E-05	0.55	0.555556	0.5	0.526316	0.100504
5	knn	9	1.00E-06	0.475	0.478261	0.55	0.511628	-0.050572
6	knn	9	1.00E-07	0.45	0.444444	0.4	0.421053	-0.100504
7	knn	9	1.00E-08	0.525	0.526316	0.5	0.512821	0.050063

nn

	model	param	a	accuracy	precision	recall	f1	mcc
0	nn	None	1.00E-01	1	1	1	1	1
1	nn	None	1.00E-02	1	1	1	1	1
2	nn	None	1.00E-03	0.625	0.619048	0.65	0.634146	0.250313
3	nn	None	1.00E-04	0.55	0.555556	0.5	0.526316	0.100504
4	nn	None	1.00E-05	0.5	0.5	0.6	0.545455	0
5	nn	None	1.00E-06	0.425	0.428571	0.45	0.439024	-0.150188
6	nn	None	1.00E-07	0.325	0.333333	0.35	0.341463	-0.350438
7	nn	None	1.00E-08	0.425	0.434783	0.5	0.465116	-0.151717

7. Synthetic Feature data – sample size 2000

rf 50

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	50	1.00E-01	1	1	1	1	1
1	rf	50	1.00E-02	0.9975	1	0.995	0.997494	0.995012
2	rf	50	1.00E-03	0.6025	0.618497	0.535	0.573727	0.206894
3	rf	50	1.00E-04	0.5375	0.536585	0.55	0.54321	0.075023
4	rf	50	1.00E-05	0.4675	0.465241	0.435	0.449612	-0.065138
5	rf	50	1.00E-06	0.495	0.494253	0.43	0.459893	-0.010086
6	rf	50	1.00E-07	0.51	0.51087	0.47	0.489583	0.020064
7	rf	50	1.00E-08	0.5125	0.512953	0.495	0.503817	0.025015

rf 100

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	100	1.00E-01	1	1	1	1	1
1	rf	100	1.00E-02	1	1	1	1	1
2	rf	100	1.00E-03	0.6175	0.634286	0.555	0.592	0.236858
3	rf	100	1.00E-04	0.4775	0.47541	0.435	0.454308	-0.045163
4	rf	100	1.00E-05	0.495	0.495	0.495	0.495	-0.01
5	rf	100	1.00E-06	0.515	0.517045	0.455	0.484043	0.030218
6	rf	100	1.00E-07	0.4875	0.487047	0.47	0.478372	-0.025015
7	rf	100	1.00E-08	0.4825	0.481865	0.465	0.473282	-0.035021

rf 150

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	150	1.00E-01	1	1	1	1	1
1	rf	150	1.00E-02	1	1	1	1	1
2	rf	150	1.00E-03	0.5975	0.596059	0.605	0.600496	0.195022
3	rf	150	1.00E-04	0.5025	0.502591	0.485	0.493639	0.005003
4	rf	150	1.00E-05	0.5	0.5	0.55	0.52381	0
5	rf	150	1.00E-06	0.53	0.533333	0.48	0.505263	0.060302
6	rf	150	1.00E-07	0.5275	0.528205	0.515	0.521519	0.055017
7	rf	150	1.00E-08	0.475	0.474227	0.46	0.467005	-0.050023

knn 1

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	1	1.00E-01	1	1	1	1	1
1	knn	1	1.00E-02	0.9975	0.995025	1	0.997506	0.995012
2	knn	1	1.00E-03	0.525	0.52451	0.535	0.529703	0.05001
3	knn	1	1.00E-04	0.54	0.539604	0.545	0.542289	0.080004
4	knn	1	1.00E-05	0.55	0.548077	0.57	0.558824	0.10008
5	knn	1	1.00E-06	0.4975	0.497143	0.435	0.464	-0.00504
6	knn	1	1.00E-07	0.5	0.5	0.56	0.528302	0
7	knn	1	1.00E-08	0.535	0.536458	0.515	0.52551	0.070056

knn 9

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	9	1.00E-01	1	1	1	1	1
1	knn	9	1.00E-02	1	1	1	1	1
2	knn	9	1.00E-03	0.5625	0.566138	0.535	0.550129	0.125189
3	knn	9	1.00E-04	0.5175	0.519553	0.465	0.490765	0.035195
4	knn	9	1.00E-05	0.5325	0.5311	0.555	0.542787	0.065066
5	knn	9	1.00E-06	0.4975	0.49635	0.34	0.403561	-0.005268
6	knn	9	1.00E-07	0.495	0.49537	0.535	0.514423	-0.010032
7	knn	9	1.00E-08	0.485	0.483696	0.445	0.463542	-0.030096

knn 5

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	5	1.00E-01	1	1	1	1	1
1	knn	5	1.00E-02	1	1	1	1	1
2	knn	5	1.00E-03	0.5425	0.544974	0.515	0.529563	0.085129
3	knn	5	1.00E-04	0.5175	0.524138	0.38	0.44058	0.036404
4	knn	5	1.00E-05	0.5425	0.544503	0.52	0.531969	0.085086
5	knn	5	1.00E-06	0.5025	0.502538	0.495	0.498741	0.005001
6	knn	5	1.00E-07	0.505	0.505952	0.425	0.461957	0.010131
7	knn	5	1.00E-08	0.5175	0.516129	0.56	0.53717	0.035127

8. Synthetic Feature data – sample size 20000

rf 50

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	50	1.00E-01	1	1	1	1	1
1	rf	50	1.00E-02	1	1	1	1	1
2	rf	50	1.00E-03	0.63525	0.647895	0.5925	0.618961	0.271494
3	rf	50	1.00E-04	0.496	0.495423	0.433	0.462113	-0.008064
4	rf	50	1.00E-05	0.49925	0.499145	0.438	0.466578	-0.001511
5	rf	50	1.00E-06	0.51625	0.518006	0.4675	0.491459	0.032656
6	rf	50	1.00E-07	0.50875	0.509621	0.4635	0.485467	0.017572
7	rf	50	1.00E-08	0.501	0.501153	0.4345	0.465453	0.002018

rf 100

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	100	1.00E-01	1	1	1	1	1
1	rf	100	1.00E-02	1	1	1	1	1
2	rf	100	1.00E-03	0.65725	0.668814	0.623	0.645094	0.31524
3	rf	100	1.00E-04	0.506	0.506243	0.4865	0.496175	0.012009
4	rf	100	1.00E-05	0.49475	0.494265	0.4525	0.472461	-0.010538
5	rf	100	1.00E-06	0.492	0.491101	0.4415	0.464982	-0.016082
6	rf	100	1.00E-07	0.5085	0.509179	0.4715	0.489616	0.017047
7	rf	100	1.00E-08	0.49825	0.498097	0.458	0.477208	-0.003511

rf 150

	model	param	a	accuracy	precision	recall	f1	mcc
0	rf	150	1.00E-01	1	1	1	1	1
1	rf	150	1.00E-02	0.99975	1	0.9995	0.99975	0.9995
2	rf	150	1.00E-03	0.6535	0.660733	0.631	0.645524	0.307311
3	rf	150	1.00E-04	0.51025	0.511147	0.47	0.489711	0.020567
4	rf	150	1.00E-05	0.5015	0.501596	0.4715	0.486082	0.003005
5	rf	150	1.00E-06	0.482	0.480851	0.452	0.465979	-0.036065
6	rf	150	1.00E-07	0.50525	0.505648	0.47	0.487173	0.010526
7	rf	150	1.00E-08	0.50525	0.505506	0.482	0.493473	0.010511

knn 1

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	1	1.00E-01	1	1	1	1	1
1	knn	1	1.00E-02	1	1	1	1	1
2	knn	1	1.00E-03	0.52425	0.525702	0.496	0.510419	0.048578
3	knn	1	1.00E-04	0.497	0.496964	0.491	0.493964	-0.006
4	knn	1	1.00E-05	0.51225	0.511841	0.5295	0.520521	0.024515
5	knn	1	1.00E-06	0.503	0.502959	0.51	0.506455	0.006001
6	knn	1	1.00E-07	0.50425	0.504335	0.4945	0.499369	0.008502
7	knn	1	1.00E-08	0.501	0.501017	0.4925	0.496722	0.002

knn 5

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	5	1.00E-01	1	1	1	1	1
1	knn	5	1.00E-02	0.99925	1	0.9985	0.999249	0.998501
2	knn	5	1.00E-03	0.5575	0.56117	0.5275	0.543814	0.115208
3	knn	5	1.00E-04	0.494	0.493852	0.482	0.487854	-0.012003
4	knn	5	1.00E-05	0.511	0.510427	0.5385	0.524088	0.022033
5	knn	5	1.00E-06	0.499	0.499007	0.5025	0.500747	-0.002
6	knn	5	1.00E-07	0.50425	0.504089	0.524	0.513851	0.008507
7	knn	5	1.00E-08	0.50025	0.500231	0.542	0.520278	0.000502

knn 9

	model	param	a	accuracy	precision	recall	f1	mcc
0	knn	9	1.00E-01	0.99975	1	0.9995	0.99975	0.9995
1	knn	9	1.00E-02	1	1	1	1	1
2	knn	9	1.00E-03	0.556	0.557613	0.542	0.549696	0.112044
3	knn	9	1.00E-04	0.50725	0.507379	0.4985	0.5029	0.014502
4	knn	9	1.00E-05	0.50575	0.505585	0.5205	0.512934	0.011505
5	knn	9	1.00E-06	0.4985	0.498562	0.52	0.509055	-0.003003
6	knn	9	1.00E-07	0.49425	0.494388	0.5065	0.50037	-0.011503
7	knn	9	1.00E-08	0.50375	0.503836	0.4925	0.498104	0.007502

nn

	model	param	a	accuracy	precision	recall	f1	mcc
0	nn	None	1.00E-01	1	1	1	1	1
1	nn	None	1.00E-02	1	1	1	1	1
2	nn	None	1.00E-03	0.6085	0.633292	0.5155	0.568357	0.220854
3	nn	None	1.00E-04	0.50425	0.504	0.5355	0.519273	0.008517
4	nn	None	1.00E-05	0.50225	0.502994	0.378	0.43163	0.004646
5	nn	None	1.00E-06	0.51125	0.510719	0.536	0.523054	0.022528
6	nn	None	1.00E-07	0.49325	0.492691	0.455	0.473096	-0.01354
7	nn	None	1.00E-08	0.4865	0.488011	0.5495	0.516933	-0.027217

Appendix IV: List of Non-Gene Physiological Features

53 Non-gene physiological features across 4 categories were included for the analysis as listed below:

Blood – 37 features

RBC

Haematocrit

hemoglobin

MCV (VGM)

MCHC (Mean corpuscular hemoglobin concentration)

MCHC% (TGMH)

WBC

Neutrophils

Eosinophils

Basophils

Monocytes

Lymphocytes

Platelets

Reticulocytes

Sodium

Potassium

Chloride

Alkaline reserve

Calcium

CPK

Urea

Creatinine

Glucose

Albumine

Proteins

Ferritin

Total Bilirubine

Phosphorus

LDH

25-OHD

PT ratio

INR

Fibrinogen

GGT

Alkaline phosphatase

AST (TGO)

ALT (TGP)

Liquid Balance – 2 Features

24 hr liquid intake

24 hr urine

Plasma Volume – 8 Features

Total Hb mass

erythroc. Volume

plasma volume

blood volume

Rel. total Hb mass

Rel. Erythroc. Volume

Rel. plasma volume

Rel. Blood volume

Vital Signs – 6 Features

Weight

Temperature

systolic PA

Diastolic PA

PA mean

Heart rate

Bibliography

[1] E. Afshinnekoo et al., “Fundamental Biological Features of Spaceflight: Advancing the Field to Enable Deep-Space Exploration,” *Cell*, Vol. 183, Nov. 2020.

[2] C. Arc-Chagnaud et al., “Evaluation of an Antioxidant and Anti-inflammatory Cocktail Against Human Hypoactivity-Induced Skeletal Muscle Deconditioning,” *Frontiers in Physiology*, Vol. 11, Feb. 2020.

[3] D. Blottner et al., “Nitrosative Redox Homeostasis and Antioxidant Response Defense in Disused Vastus Lateralis Muscle in Long-Term Bedrest (Toulouse Cocktail Study),” *Antioxidants*, Vol. 10, Mar. 2021.

[4] T. Liu, G. Melkus, T. Ramsay, A. Sheikh, O. Laneuville, and G. Trudel, “Bone Marrow Reconversion With Reambulation: A Prospective Clinical Trial,” *Investigative Radiology*, Vol. 56, Apr. 2021.

[5] K. Austermann et al., “Antioxidant Supplementation Does Not Affect Bone Turnover Markers During 60 Days of 6° Head-Down Tilt Bed Rest: Results from an Exploratory Randomized Controlled Trial,” *The Journal of Nutrition,* Vol. 151, Issue 6, Jun. 2021.

[6] K. Austermann et al., “Effects of Antioxidant Supplementation on Bone Mineral Density, Bone Mineral Content and Bone Structure in Healthy Men During 60 Days of 6° Head-Down Tilt Bed Rest: Results from a Randomised Controlled Trial,” *Nutrition Bulletin*, Vol. 48, Issue 2, Jun. 2023.

[7] B. Hoffmann, P. Dehkordi, F.K. Khavar, N. Goswami, A. P. Blaber, K. Tavakolian, “Mechanical Deconditioning of the Heart Due to Long-Term Bed Rest as Observed on Seismocardiogram Morphology,” *npj Microgravity,* Jul. 2022.

[8] S. Solbiati et al., “Analysis of Changes in Cardiac Circadian Rhythms of RR and QT Induced by a 60-Day Head-Down Bed Rest With and Without Nutritional Countermeasure,” *European Journal of Applied Physiology*, Vol. 120, Jul. 2020.

[9] K. Culliton, H. Louati, O. Laneuville, T. Ramsay, and G. Trudel, “Six Degrees Head-Down Tilt Bed Rest Caused Low-Grade Hemolysis: A Prospective Randomized Clinical Trial,” *npj Microgravity,* Feb. 2021.

[10] P. Barbacini et al., “Effects of Omega-3 and Antioxidant Cocktail Supplement on Prolonged Bed Rest: Results from Serum Proteome and Sphingolipids Analysis,” *Cells,* Jul. 2021.

[11] W. V. Trim et al., “Impact of Long-Term Physical Inactivity on Adipose Tissue Immunometabolism,” *The Journal of Clinical Endocrinology & Metabolism,* Vol. 107, Jan. 2022.

[12] K. Brauns, A. Friedl-Werner, H.C. Gunga, and A.C. Stahn, “Effects of Two Months of Bed Rest and Antioxidant Supplementation on Attentional Processing,” *Cortex,* Vol. 141, Apr. 2021.

[13] P. Jacob, J. Bonnefoy, S. Ghislin and J.P. Frippiat, “Long-Duration Head-Down Tilt Bed Rest Confirms the Relevance of the Neutrophil to Lymphocyte Ratio and Suggests Coupling It with the Platelet to Lymphocyte Ratio to Monitor the Immune Health of Astronauts,” *Frontiers in Immunology,* Vol. 13, Oct. 2022.

[14] M. Kermorgant et al., “The Effects of Antioxidant Cocktail on Ophthalmological Changes Induced by a 60-Day Head-Down Bed Rest in a Randomized Trial,” *Life,* Vol. 14, Dec. 2024.

[15] T. Louwies et al., “Retinal Blood Vessel Diameter Changes with 60-Day Head-Down Bedrest Are Unaffected by Antioxidant Nutritional Cocktail,” *npj Microgravity,* Nov. 2024.

[16] P. Sundblad, O. Orlov, O. Angerer, I. Larina, and R. Cromwell, “Standardization of Bed Rest Studies in the Spaceflight Context,” *Journal of Applied Physiology,* Vol. 121, Jul. 2016.

[17] D. E. Watenpaugh, “Analogs of Microgravity: Head-Down Tilt and Water Immersion,” *Journal of Applied Physiology,* Vol. 120, Apr. 2016.

[18] F. Ferranti, M. Del Bianco, and C. Pacelli, “Advantages and Limitations of Current Microgravity Platforms for Space Biology Research,” *Applied Sciences,* Vol. 11, Dec. 2020.

[19] Y. Hao et al., “Integrating Bioinformatic Strategies in Spatial Life Science Research,” *Briefings in Bioinformatics*, Vol. 23, Issue 6, Nov. 2022.

[20] S. S. Ahmed et al., “Systematic Review of the Effectiveness of Standalone Passive Countermeasures on Microgravity-Induced Physiologic Deconditioning,” *npj Microgravity,* Apr. 2024.

[21] J. V. Meck, S. A. Dreyer, and L. E. Warren, “Long-Duration Head-Down Bed Rest: Project Overview, Vital Signs, and Fluid Balance,” *Aviation, space, and environmental medicine,* Jun. 2009.

[22] D. Stratis, G. Trudel, L. Rocheleau, M. Pelchat, and O. Laneuville, “The Characteristic Response of the Human Leukocyte Transcriptome to 60 Days of Bed Rest and to Reambulation,” *Medicine and science in sports and exercise,* Oct. 2022.

[23] P. Domingos, “A Few Useful Things to Know About Machine Learning,” *Commun. ACM,* Vol. 55, pp. 78–87, Oct. 2012.

[24] L. Breiman, “Random Forests,” *Machine Learning,* Vol. 45, pp. 5–32, Oct. 2001.

[25] Q. Zhu, “On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset,” Pattern Recognition Letters, Vol. 136, pp. 71–80, 2020.

[26] S. Varma and R. Simon, “Bias in Error Estimation When Using Cross-Validation for Model Selection,” *BMC Bioinformatics,* Vol. 7, p. 91, Feb. 2006.

[27] M. Kuhn and K. Johnson, “Applied Predictive Modeling,” *Springer*, 2013.

[28] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory,* Vol. 13, pp. 21–27, Jan. 1967.

[29] R. Duda, “Pattern Classification,” 2nd ed. 2000.

[30] A. Liaw and M. Wiener, “Classification and Regression by RandomForest,” *R News,* Vol. 2, pp. 18–22, Dec. 2002.

[31] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning” *MIT Press,* Oct. 2016.

[32] M. A. Nielsen, “Neural Networks and Deep Learning,” *Determination Press,* 2018.

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research,* Vol. 16, pp. 321–357, Jun. 2002.

[34] B. Frénay and M. Verleysen, “Classification in the Presence of Label Noise: A Survey,” *IEEE Transactions on Neural Networks and Learning Systems,* Vol. 25, pp. 845–869, May. 2014.

[35] C. Shorten and T. M. Khoshgoftaar, “A Survey on Image Data Augmentation for Deep Learning,” *Journal of Big Data,* Vol. 6, Jul. 2019.

[36] G. B. Gloor, J. M. Macklaim, V. P. Glahn, and J. J. Egozcue, “Microbiome Datasets Are Compositional: And This Is Not Optional,” *Frontiers in Microbiology,* Vol. 8, Nov. 2017.

[37] J. Aitchison, “The Statistical Analysis of Compositional Data,” *Journal of the Royal Statistical Society,* Vol. 44, 1982.

[38] T. P. Quinn, I. Erb, G. Gloor, C. Notredame, M. F. Richardson, and T. M. Crowley, "A Field Guide for the Compositional Analysis of Any-Omics Data," *GigaScience,* Vol. 8, Sept. 2019.

[39] L. Devroye, L. Györfi, and G. Lugosi, "A Probabilistic Theory of Pattern Recognition," *Springer,* Jan. 1996.

[40] C. M. Bishop, "Pattern Recognition and Machine Learning," *Springer,* Oct. 2007.

[41] M. I. Love, S. Anders, V. Kim, and W. Huber, "RNA-Seq workflow: gene-level exploratory analysis and differential expression," *F1000Research,* Vol. 4, Oct. 2015.

[42] M. I. Love, W. Huber, and S. Anders, "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2," *Genome Biology,* Vol. 15, Dec. 2014.

[43] B. Iglewicz and D. C. Hoaglin, "How to Detect and Handle Outliers," *ASQC Quality Press*, Vol. 6, 1993.

[44] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research,* Vol. 15, pp. 3133–3181, Oct. 2014.

[45] Z. H. Zhou, "Ensemble Methods: Foundations and Algorithms," *CRC Press,* Jun. 2012.

[46] S. Raudys and A. K. Jain, "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.,* Vol. 13, pp. 252–264, 1991.

[47] S. Mukherjee et al., "Estimating Dataset Size Requirements for Classifying DNA Microarray Data," *Journal of Computational Biology: a Journal of Computational Molecular Cell Biology,* Vol. 10, pp. 119–142, Feb. 2003.

[48] P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern Classification with Missing Data: A Review," *Neural Computing and Applications,* Vol. 19, pp. 263–282, Mar. 2010.

[49] X. W. Mao et al., "Spaceflight Environment Induces Mitochondrial Oxidative Damage in Ocular Tissue," *Radiation Research,* Vol. 180, Sept. 2013.

[50] E. Waisberg, et al., "Challenges of Artificial Intelligence in Space Medicine," *Space: Science & Technology,* Oct. 2022.

[51] R. T. Scott, et al., "Biomonitoring and precision health in deep space supported by artificial intelligence," *Nature Machine Intelligence,* Vol. 5, pp. 196–207, Mar. 2023.

[52] D. Fernandes, J. M. Macklaim, T. G. Linn, G. Reid, and G. B. Gloor, "ANOVA-like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq," *PLoS ONE,* Vol. 9, Jul. 2013.

[53] S. D. Childress, T. C Williams, and D. R. Francisco, "NASA Space Flight Human-System Standard: enabling human spaceflight missions by supporting astronaut health, safety, and performance," *npj Microgravity,* Vol. 9, Apr. 2023.

[54] S. M. Smith et al., "Effects of Artificial Gravity During Bed Rest on Bone Metabolism in Humans," *Journal of Applied Physiology,* Dec. 2008.

[55] R. Hargens and L. Vico, "Long-Duration Bed Rest as an Analog to Microgravity," *Journal of Applied Physiology,* Vol. 120, pp. 891–903, Feb. 2016.

[56] M. Knovich, J. Storey, L. Coffman, S. Torti, and F. Torti, "Ferritin for the Clinician," *Blood reviews,* Vol. 23. pp. 95–104. Nov. 2008.

[57] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal Machince Learning Research,* Vol. 12, pp. 2825–2830, Oct. 2011.

[58] B. L. Dickerson, R. Sowinski, R. B. Kreider, and G. Wu, "Impacts of microgravity on amino acid metabolism during spaceflight," *Experimental Biology and Medicine*, Vol. 248, pp. 380–393, Feb. 2023.

[59] W. Trim et al., "The impact of physical inactivity on glucose homeostasis when diet is adjusted to maintain energy balance in healthy, young males," *Clinical Nutrition*, Vol. 42, pp. 532–540, Apr. 2023.

[60] N. F. Shur et al., "Human adaptation to immobilization: Novel insights of impacts on glucose disposal and fuel utilization," *Journal of Cachexia, Sarcopenia and Muscle*, Vol. 13, pp. 2999–3013, Sep. 2022.

[61] L. Peter, M. Chung, Z. Ren, D. B. Mair, and D.H. Kim, "Factors mediating spaceflight-induced skeletal muscle atrophy," *American Journal of Physiology-Cell Physiology,* Vol. 322, Feb. 2022.

[62] A. Salon et al., "Retinal venular vessel diameters are smaller during ten days of bed rest," *Scientific Reports,* Vol. 13, Nov. 2023.

[63] I. Hussain et al., "Cardiovascular effects of long-duration space flight," *Health Science Reports,* Aug. 2024.

[64] C. Ludtka, J. Silberman, E. Moore, and J. B. Allen, “Macrophages in microgravity: the impact of space on immune cells,” *npj Microgravity,* Vol. 7, Mar. 2021.

[65] S. Tauber et al., “Signal Transduction in Primary Human T Lymphocytes in Altered Gravity During Parabolic Flight and Clinostat Experiments,” *Cellular Physiology and Biochemistry,* Feb. 2015.

[66] J. K. Smith, “The effects of microgravity on cellular elements of the central nervous system - astrocytes, microglia, neurons, oligodendrocytes, and oligodendrocyte precursor cells,” *International Journal of Frontline Research and Reviews,* Vol. 2, pp. 9–18, Feb. 2024.

[67] W. Zhang et al.,68“Structural and functional insights into the lipid regulation of human anion exchanger 2,” *Nature Communications,* Vol. 15, Jul. 2024.

[68] K. Paulsen et al., “Severe disruption of the cytoskeleton and immunologically relevant surface molecules in a human macrophageal cell line in microgravity—Results of an in vitro experiment on board of the Shenzhou-8 space mission,” *Acta Astronautica,* Vol. 94, 2014.

[69] J. Casey, S. Grinstein, and J. Orlowski, “Sensors and regulators of intracellular pH,” *Nature reviews Molecular cell biology,* 2009.

[70] G. Rea et al., “Microgravity-driven remodeling of the proteome reveals insights into molecular mechanisms and signal networks involved in response to the space flight environment,” *Journal of Proteomics,* Vol. 137, pp. 3–18, Mar. 2016.

[71] K. B. Sheehan, K. McInnerney, B. Purevdorj-Gage, S. D. Altenburg, and L. E. Hyman, “Yeast genomic expression patterns in response to low-shear modeled microgravity,” *BMC Genomics,* Vol. 8, Jan. 2007.

[72] P. A. Plett, R. Abonour, S. M. Frankovitz, and C. M. Orschell, “Impact of modeled microgravity on migration, differentiation, and cell cycle control of primitive human hematopoietic progenitor cells,” Experimental Hematology, Vol. 32, pp. 773–781, Aug. 2004.

[73] M. A. Travis and D. Sheppard, “TGF-β Activation and Function in Immunity,” *Annual Review of Immunology,* Vol. 32, pp. 51–82, Mar. 2014.

[74] S. Avram, S. Shaposhnikov, C. Buiu, and M. Mernea, “Chondroitin Sulfate Proteoglycans: Structure-Function Relationship with Implication in Neural Development and Brain Disorders,” *BioMed Research International,* Vol. 2014, pp. 1–11, Apr. 2014.

[75] R. V. Iozzo and L. Schaefer, “Proteoglycan form and function: A comprehensive nomenclature of proteoglycans,” *Matrix Biology,* Vol. 42, pp. 11–55, Mar. 2015.

[76] Z. Che, S. Purushotham, K. Cho, D. Sontag and Y. Liu, “Recurrent Neural Networks for Multivariate Time Series with Missing Values,” *Scientific Reports,* Vol. 8, Apr. 2018.

[77] A. Kassambara, “Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning,” 1st ed., *STHDA,* 2017.

[78] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, “Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance.” *Technologies,* Vol. 9, Jul. 2021.

[79] S. M. Parry and Z. A. Puthucheary, “The impact of extended bed rest on the musculoskeletal system in the critical care environment,” *Extreme Physiology & Medicine,* Vol. 4, Oct. 2015.

[80] S. Marchal, A. Choukér, J. Bereiter-Hahn, A. Kraus, D. Grimm and M. Krüger “Challenges for the human immune system after leaving Earth,” *npj Microgravity,* Vol. 10, Nov. 2024.

[81] S. M. Smith, S. R. Zwart, G. Bock, B. L. Rice, and J. E. Davis-Street, “The Nutritional Status of Astronauts Is Altered after Long-Term Space Flight Aboard the International Space Station,” *The Journal of Nutrition,* Vol. 135, Issue 3, pp. 437–443, Mar. 2005.

[82] T. Haider et al., “Effects of long-term head-down-tilt bed rest and different training regimes on the coagulation system of healthy men,” *Physiological reports,* 2013.

[83] G. Austin and T. Korem, “Compositional transformations can reasonably introduce phenotype-associated values into sparse features,” *mSystems,* Vol. 10, 2025.

[84] S. Yelmen et al., “Creating artificial human genomes using generative neural networks,” *PLOS Genetics,* Vol. 17, Feb. 2021.

[85] A. Budhkar, Q. Song, J. Su, and X. Zhang, “Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics,” *Computational and Structural Biotechnology Journal,* Vol. 27, pp. 346–359, Dec. 2024.

[86] T. C. Liu, P.N. Kalugin, J. L. Wilding, W. F. Bodmer, “GMMchi: gene expression clustering using Gaussian mixture modeling,” *BMC Bioinformatics,* Vol. 23, Nov. 2022.

[87] D. Krstajic, L. J. Buturovic, D. E. Leahy, S. Thomas, “Cross-validation pitfalls when selecting and assessing regression and classification models,” *Journal of Cheminformatics,* Vol. 6, Mar. 2014.

[88] S. B. Kotsiantis, “Supervised machine learning: A review of classification techniques,” *Emerging Artificial Intelligence Applications in Computer Engineering,* Vol. 160, pp. 3–24, 2007.

[89] R. Dinga, L. Schmaal, B. Penninx, D. J. Veltman, A. F. Marquand, “Controlling for effects of confounding variables on machine learning predictions,” *BioRxiv,* Aug. 2020.

[90] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and Intelligent Laboratory Systems,* Vol. 2, pp. 37–52, Aug. 1987.

[91] P. R. Sarma, “Red Cell Indices,” *Clinical Methods: The History, Physical, and Laboratory Examinations,* 3rd edition, Chapter 152, 1990.

[92] T. Ju and R.D. Cummings, “A unique molecular chaperone Cosmc required for activity of the mammalian core 1 β3-galactosyltransferase,” *Proceedings of the National Academy of Sciences of the United States of America,* 2002.

[93] G. Eraslan, Z. Avsec, J. Gagneur, and F. Theis, “Deep learning: new computational modelling techniques for genomics,” *Nature Reviews Genetics,* Vol. 20, pp. 389–403, 2019.

[94] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction,” 2nd ed., *Springer,* 2009.

[95] F. Jiang, “Artificial intelligence in healthcare: past, present and future,” *Stroke and Vascular Neurology,* Vol. 2, no. 4, pp. 230–243, Jun. 2017.

[96] S. Lundberg and S. Lee, “A Unified Approach to Interpreting Model Predictions,” *Proceedings of the 31st International Conference on Neural Information Processing Systems,* 2017.

[97] M. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” Nature Reviews Genetics, Vol. 16, pp. 321–332, Jun. 2015.

[98] T. Osabe, K. Shimizu, and K. Kadota, “Differential expression analysis using a model-based gene clustering algorithm for RNA-seq data,” *BMC Bioinformatics,* Vol. 22, Oct. 2021.

[99] S. M. Smith, J. M. Krauhs, C. S. Leach, “Regulation of Body Fluid Volume and Electrolyte Concentrations in Spaceflight,” *Advances in Space Biology and Medicine,* Vol. 6, pp. 123–165, 1997.

[100] S. Uddin, I. Haque, H. Lu, M.A. Moni, and E. Gide, “Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction,” *Scientific Reports,* Vol. 12, Apr. 2022.

[101] L. Li, T. A. Darden, C.R. Weinberg, A.J. Levine, and L.G. Pedersen, "Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method," *Comb Chem High Throughput Screen,* Dec. 2001.

[102] S. A. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule," in *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. SMC-6, no. 4, pp. 325–327, Apr. 1976.

[103] P. Cunningham and S. J. Delany, "k-Nearest neighbour classifiers," *ACM Computing Surveys,* Vol. 54, Apr. 2007.

[104] H. Alsagri and M. Ykhlef, "Quantifying Feature Importance for Detecting Depression using Random Forest," *International Journal of Advanced Computer Science and Applications,* Vol. 11, Jan. 2020.

[105] T. M. Oshiro, P.S Perez, and J. A Baranauskas, "How many trees in a random forest?" *Machine Learning and Data Mining in Pattern Recognition (MLDM), Springer,* pp. 154–168, 2012.

[106] P. Probst and A. L Boulesteix, "To tune or not to tune the number of trees in random forest," *Journal of Machine Learning Research,* Vol. 18, pp. 1–18, 2018.

[107] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning,* Vol. 63, no. 1, pp. 3–42, Mar. 2006.

[108] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation." *Journal of Machine Learning Technologies,* Vol. 2, pp. 37–63, 2011.

[109] J. J. Egozcue, V. Pawlowsky-Glahn, G. Figueras, and C. Vidal, "Isometric Logratio Transformations for Compositional Data Analysis," *Mathematical Geology,* Vol. 35. pp. 279–300, Apr. 2003.