APC524 - Final Project Proposal

Shelby Fulton Boaz Yoo

September 2025

Scope

Female athletes are a heavily underserved population in terms of sports science and understanding how to optimize performance and avoid injury. Less than 6% of sports science research has been conducted exclusively on female athletes [2], leaving this subject as a mystery to all. Studies also suggest that there are correlations between certain injuries and the athlete's menstrual cycle phase [4]. There are also other issues concerning the menstrual cycle and athletic performance such as psychological changes throughout the cycle that impact athletes' mental health and performance [3] and also RED-S (Relative Energy Deficiency in Sport) syndrome [1]. That is why we are in the process of building an application to help female athletes understand their menstrual cycle and how to work with their body, instead of against it, in order to achieve their athletic goals. The application takes in both menstrual cycle symptoms that are manually logged each day along with training data collected by wearables.

The objective of this project is to build a pipeline that allows for collecting and streaming health data from wearable devices into a machine learning pipeline that processes the data and trains and evaluates a model that can propose the most adequate nutritional, recovery, and performance recommendations based on the health data. Such a tool will help advancements in research in women's health and athletics, as well as facilitate the creation of products that female athletes can use to optimize their health and performance.

Content

To accomplish this, our project will have the following components:

1. **Synthetic data generator (SDG)** to create simulation data for both wearable training data and menstrual cycle symptoms. Although the ultimate goal is the directly stream health data from wearable devices or through Apple HealthKit to a machine learning pipeline, we will start off with synthetic generation of the which can be more readily tested programmatically within the scope of this course.
2. **Machine learning pipeline** which will include the:
   (a) **Data cleaning + preprocessing module (DPM)** to reformat the data/engineer features usable and adequate for the machine learning model.
   (b) **Machine learning model training** to use real + synthetic data that ideally can predict symptoms and athletic performance based on previous data to a certain accuracy threshold as well as prescribe nutritional, recovery, and performance recommendations.
   (c) **Evaluation** of the ML model.
3. **API** endpoint to send the wearable data (which we will simulate with the **SDG**) to the **DPM** in the **machine learning pipeline**.

Implementation

The project will be implemented in **Python**. Various computing tasks (mainly in the **machine learning pipeline**) will be implemented using **NumPy**, **pandas**, **SciPy**, **scikit-learn**, and **Tensorflow**. To interact with GPUs during, we will use **CUDA**.

We will do version control with **git** and share/host on **GitHub**.

We will test the **SDG** to make sure realistic random data is correctly generated. We will test that the **DPM** as well as the ML model itself accepts inputs/outputs with of the correct shape/dimensions. We will test that requests made to the **API** correctly is sent to the **DPM/machine learning pipeline**.

We will primarily test these with **functional** and **property based testing** at both **unit** and **integration test** scopes. These tests will be run with the **pytest** framework.

To make this project distributable and usable by others [5], we will package our project. We will use **uv** for building and publishing the package and **GitHub Actions** for **continuous integration**.

References

[1] H. E. Cabre, S. R. Moore, A. E. Smith-Ryan, and A. C. Hackney. Relative energy deficiency in sport (red-s): Scientific, clinical, and practical implications for the female athlete. *Deutsche Zeitschrift für Sportmedizin*, 73(7):225–234, November 2022.

[2] Emma S. Cowley, Alyssa A. Olenick, Kelly L. McNulty, and Emma Z. Ross. "invisible sportswomen": The sex data gap in sport and exercise science research. *Women in Sport and Physical Activity Journal*, 29(2):146–151, October 2021.

[3] Scott Gilmour. The psychological impact of the menstrual cycle on athletic performance. https://sportrxiv.org/index.php/server/preprint/view/503, January 2025. Preprint, SportRxiv. Accessed: 2025-09-09.

[4] Candice MacMillan, Benita Olivier, Carel Viljoen, Dina Christa Janse van Rensburg, and Nicola Sewry. The association between menstrual cycle phase, menstrual irregularities, contraceptive use and musculoskeletal injury among female athletes: A scoping review. *Sports Medicine*, 54(10):2515–2530, August 2024.

[5] Henry Schreiner, Romain Teyssier, Gabriel Perez-Giz, and Troy Comi. Software engineering for scientific computing. https://github.com/se-for-sci/se-for-sci.github.io, 2025. Accessed: 2025-09-09.