DRAFT – Main Title Here – DRAFT

A Machine Learning Approach to Correlating Physiological Parameters with Rheological Features

A Manning Smith

University of Delaware
Bioinformatics and Computation Science

July 12, 2025

Abstract

Abstract goes here

Contents

1 Introduction

Intro section here.

2 Methods

The data sampling methods.. Need from Sean....

This work was all performed using Python version 3.12.3. The utilized packages are outlined in the supplied requirements.txt found in the supplemental resources' section for reproducibility. Various machine learning methods were utilized in alignment with a dataset comprised of a small sample size. The Scikit-learn package, version 1.6.1 was accessed for all the machine learning techniques. The accuracy of the models was validated through K-Fold Cross Validation. The utilized methods were produced with a constant random seed of 1743, the year University of Delaware was founded.

To start with data processing it was important to address missing values in the rheological dataset due to the already limited sample size. K-nearest neighbors (KNN) imputation was used for data imputation to fill out the missing values forming a complete set of rheological data. The KNN algorithm identifies the k most similar complete observations based on Euclidean distance across all available features, then imputes missing values using the weighted average of these neighbors' corresponding values. The chosen value of $k = 3$ captures local data structure while maintaining robustness against outliers. This approach is well-suited for rheological data, where the parameters often exhibit strong intercorrelations due to their shared dependence.

Secondly, the physiological variables were standardized using sklearn's StandardScaler, which transforms each feature to have zero mean and unit variance according to:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where x_{ij} represents the original value for sample i and feature j, μ_j is the sample mean, and σ_j is the sample standard deviation for feature j.

Standardization was essential for the PCA, as PCA is sensitive to the relative scales of input variables. Without standardization, variables with larger natural scales (e.g., cholesterol levels in mg/dL) would dominate the principal components over variables with smaller scales (e.g., hematocrit percentages), leading to biased dimensionality reduction that reflects measurement units rather than biological relationships.

Previous research (cite paper) highlighted viability in the physiological variables, however further dimensional reduction may prove further viability preserving the maximum variance. PCA transforms the original feature space into a new coordinate system where the axes (principal components) are linear combinations of the original variables, ordered by the amount of variance they explain.

The mathematical foundation involves eigendecomposition of the covariance matrix:

$$C = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

where $\mathbf{X}$ is the standardized data matrix. The principal components are the eigenvectors of $\mathbf{C}$, and the explained variance for each component corresponds to its eigenvalue. For the purposes of this research the following 13 features were selected: 'HCT', 'FIB', 'CHOL', 'TRIG', 'HDL', 'LDL', 'WBC', 'RBC', 'HEM', 'MCV', 'MCH', 'MCHC'.

The analysis focused primarily on the first 8 components, which captured the majority of physiological variation while maintaining sufficient sample-to-feature ratios for robust statistical analysis. Principal component loadings were examined to identify which physiological variables contributed most strongly to each component, enabling biological interpretation of the reduced feature space.

In starting with the predictability of the PCs Pearson product-moment correlation coefficients were calculated to assess linear relationships between principal components and rheological parameters. Pearson correlation was selected as the method for showing viability before advancing to further machine learning techniques. Both principal components and rheological measurements represent continuous variables measured on interval scales, satisfying the data type requirements for parametric correlation analysis. Principal components are linear combinations of the original physiological variables, making linear correlation analysis theoretically appropriate for detecting associations with rheological behavior.

Upon showing viability through Pearson correlation, Gaussian Process Regression was implemented to model the complex relationships between PCs derived from physiological parameters and individual rheological parameters. GPR was selected specifically for its ability to provide uncertainty quantification, handle non-linear relationships, and perform reasonably well with small datasets. The GPR framework assumes that the target function $f(x)$ follows a Gaussian process defined by $f(x) \approx GP(m(x), k(x, x'))$, where $m(x)$ represents the mean function (set to zero) and $k(x, x')$ represents the covariance function that determines similarity between inputs.

The kernel selection strategy employed a composite kernel combining Radial Basis Function and White Noise components according to $k(x, x') = k_{RBF}(x, x') + k_{white}(x, x')$, where bounded constraints on kernel parameters prevented pathological solutions such as memorization of training points or reduction to pure noise models. The RBF kernel captures smooth, non-linear relationships through the mathematical formulation $k_{RBF}(x, x') = \sigma_f^2 \exp\left(\frac{-|x-x'|^2}{2l^2}\right)$, where σ_f^2 represents the signal variance controlling the overall amplitude of function variations, l denotes the length scale parameter determining how quickly correlations decay with distance between inputs, and $|x - x'|^2$ represents the squared Euclidean distance between data points in the principal component space. Hyperparameter bounds set to length_scale_bounds= $(1 \times e^{-2}, 1 \times e^{2})$ and noise_level_bounds= $(1 \times e^{-5}, 1 \times e^{0})$ to prevent degenerate solutions while maintaining sufficient flexibility to capture meaningful physiological-rheological relationships.

For validating the implemented GPR method, K-fold Cross-Validation was utilized. $k = 5$ folds was selected to balance the need for adequate training data with robust performance assessment given the limited sample size of $n = 22$. This methodology partitions the dataset into five approximately equal segments, with each segment serving as a validation set once while the remaining segments form the training set, repeating this process five times with performance metrics calculated for each iteration. The approach was configured with $n_splits = 5$ creating folds containing 4-5 samples each for testing, shuffle=True to randomize sample assignment and prevent systematic bias from potential temporal ordering in data collection or donor recruitment patterns, and our $random_state = 1743$ to ensure reproducible fold assignments across analysis runs.

K-fold cross-validation helped in for identifying various challenges associated with applying GPR to high-dimensional data with small sample sizes. The cross-validation framework illuminated the severity of overfitting issues inherent in GPR applications to datasets where the sample-to-feature ratio approaches problematic thresholds, as evidenced by the substantial gap between training performance and validation performance across all folds. This diagnostic capability proved instrumental in understanding the limitations of the modeling approach and informed subsequent methodological decisions regarding regularization strategies, with the consistent negative performance across folds providing clear evidence that the challenges were not attributable to specific data partitions or outlier influences but rather represented fundamental mismatches between model complexity and available training data. The stable, reproducible results obtained through the k-fold framework thus served both to rigorously evaluate model performance and to provide critical insights into the broader challenges of machine learning applications in small-sample biological datasets.

Given the limitations imposed by the small sample size relative to the high-dimensional feature space, synthetic data generation was employed as a strategy to augment the training dataset and improve machine learning model performance. Two complementary approaches were implemented to generate realistic synthetic samples while preserving the underlying statistical relationships in the original dataset.

The fundamental challenge in synthetic data generation for biological systems lies in maintaining both the marginal distributions of individual variables and the complex correlation structure between physiological and rheological parameters. Simple parametric approaches often fail to capture the non-linear dependencies present in biological data, while purely random sampling may generate physiologically implausible combinations of values.

The first approach implemented a KNN synthetic data generator that leverages local similarity structures within the original dataset, operating on the principle that realistic synthetic samples should lie within the convex hull defined by existing observations, particularly near regions of high data density where the likelihood of encountering similar physiological profiles in real populations is highest. The KNN synthetic generation algorithm proceeds through a systematic four-step process beginning with seed point selection, where for each synthetic sample a seed point is randomly selected from the original dataset to serve as the basis for generation, ensuring that all synthetic samples are anchored to actual observed physiological profiles. Neighborhood identification follows, where the $k = 3$ nearest neighbors to the seed point are identified using Euclidean distance in the standardized feature space, with our chosen k to balance between capturing local data structure while maintaining sufficient diversity to prevent over-reliance on any single neighboring observation.

The second approach employed a Gaussian copula methodology to capture and reproduce the full joint distribution of physiological and rheological variables. Copula-based methods separate the modeling of marginal distributions from the dependence structure, allowing for flexible representation of complex multivariate relationships.

Variable transformation: Variables are transformed to uniform marginals using rank-based empirical cumulative distribution functions:

$$U_{ij} = (rank(X_{ij}) + 1)/(n + 1)$$

n is the sample size.

New samples are drawn from the multivariate normal distribution and inverse-transformed back to their original scales. For enhancing the relationships between the variables KNN imputation methods were used to preserve the original data structure utilizing information from similar cases.

Model performance was evaluated using multiple complementary metrics to provide a comprehensive assessment of predictive capability. The Cross-Validated R^2 Score served as the primary performance metric. The R^2_{CV} represents the proportion of variance explained in the response variable. This metric provides an unbiased estimate of the models explanatory power. Values closer to 1 suggest better predictive performance. The R^2_{train} was calculated on the training dataset used in fitting the model. This metric helps as a baseline when comparing the cross-validation. Other metrics such as Mean Squared Error Cross-Validation (MSE_{CV}) quantifying the prediction error magnitude, helping to show predictive accuracy on the unseen data. Then the overfitting assessment serves as an evaluation of assessing the models' generalization capability, calculated as the combination of the and R^2_{CV} and MSE_{CV}.

While formal multiple comparison corrections were not applied due to the exploratory nature of the analysis, results were interpreted with appropriate caution. The focus was placed on identifying patterns and effect sizes rather than definitive statistical conclusions, with emphasis on replication and validation in future larger datasets. The hierarchical nature of the analysis (PCA followed by correlation analysis) was considered when interpreting statistical significance, recognizing that the reduced feature space may concentrate signal while also potentially introducing bias in significance testing. This methodological framework provides a comprehensive approach to analyzing the relationship between physiological blood parameters and rheological behavior while acknowledging and addressing the inherent challenges of working with limited sample sizes in complex biological systems.

3 Results and Discussion

Results and discussion section here.

4 Future Directions

Future directions section here.

5 Conclusion

Conclusion section here.

6 Acknowledgments

- Sean Farrington - Norman J. Wagner

7 Resources

Resources section here.

8 Acronyms

- GPR: Gaussian Process Regression
- PC: Principal Component
- ML: Machine Learning
- PCA: Principal Component Analysis
- CV: Cross-Validation
- KFCV: K-Fold Cross-Validation
- MSE: Mean Squared Error
- RMSE: Root Mean Squared Error
- MAE: Mean Absolute Error
- R^2: Coefficient of Determination
- CI: Confidence Interval
- PCs: Principal Components

A Supplementary Data

[Include supplementary figures, tables, or data here]