# Identification of Non-invasive Cytokine Biomarkers for Polycystic Ovary Syndrome Using **Supervised Machine Learning**

**Daniela Perry[1,2], Tathagata Dasgupta[2], Joseph Dexter[2], Sarah Field[3], Michele Cummings[3], Vinay Sharma[3], Nadia Gopichandran[3], Ellis Baskind[3], Nicholas Orsi[3], Jeremy Gunawardena[2]**
[1]Cornell University, Department of Biological Statistics and Computational Biology [2]Harvard Medical School, Department of Systems Biology [3]University of Leeds Institute of Cancer and Pathology

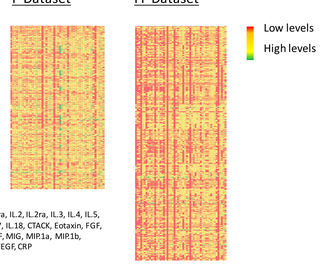
## Abstract

Polycystic ovary syndrome (PCOS) is a common endocrine disorder that affects up to 20% of women, however diagnosis is commonly unreliable and un-quantitative. Here we use supervised machine learning and measurements of 51 cytokines from a large cohort of patients to identify a low-dimensional set of potential biomarkers for diagnosis of PCOS. Both whole blood and individual follicular fluid (FF) aspirates were collected women during pre-intracytoplasmic sperm injection with in vitro fertilization (ICSI/IVF) oocyte retrieval and linked with patients' PCOS status as diagnosed by the Rotterdam criteria (n = 69 PCOS, n = 222 non-PCOS). We trained a binary support vector machine (SVM) using a random subset of patient data to determine cytokine profile associated with PCOS. Our resultant model includes 3 variables and is 76% accurate. This provides insight into the immunological basis of PCOS and may define a potential non-invasive quantitative strategy for diagnosis.

## Introduction

PCOS is an endocrine disorder that affects up to 20% of women. It is diagnosed using the Rotterdam criteria, which are as follows:


With these diagnosis criteria in mind, one thing my projects aims to accomplish is to provide a different measurement of PCOS using concrete levels of cytokines. This study involved taking samples from 291 women by first exposing them to long-protocol ovarian hyperstimulation, which is a technique used to induce ovulation by multiple ovarian follicles. Then samples are taken from each woman's blood plasma and follicular fluid. To note, the reason why there is approximately double the amount of follicular fluid samples is that for this set, each ovary is sampled, yielding roughly double the number of samples. Then, fifty-one whole blood and FF cytokines were measured by fluid-phase multiplex cytometric immunoassay (the resultant dataset is pictured below). The different cytokines were detected using different antibodies, which can be quite an expensive and lengthy test. So, another goal of this projec is to reduce the number of cytokines, or features, needed to predict PCOS in patients.




Dataset visualization. These two figures represent the plasma (P) and follicular fluid (FF) datasets which are 291x51 and 530x51 respectively.

A comprehensive list of the 48 cytokines measured is: IL.1a, IL.1b, IL.1ra, IL.2, IL.2ra, IL.3, IL.4, IL.5, IL.6, IL.7, IL.8, IL.9, IL.10, IL.12..p40., IL.12..p70., IL.13, IL.15, IL.16, IL.17, IL.18, CTACK, Eotaxin, FGF, G.CSF, GM.CSF, GRO.a, IFN.a, IFN.g, IP.10, LIF, MCP.1, MCP.3, M.CSF, MIF, MIG, MIP.1a, MIP.1b, b.NGF, PDGF, RANTES, SCF, SDF.1a, TGF.b, TGF.b, TNF.a, TNF.b, TRAIL, VEGF, CRP

## Definitions & Equations

**Cytokines:** Cytokines are small secreted proteins released by cells have a specific effect on the interactions and communications between cells.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Accuracy} = \frac{TN}{TN + FP}$$

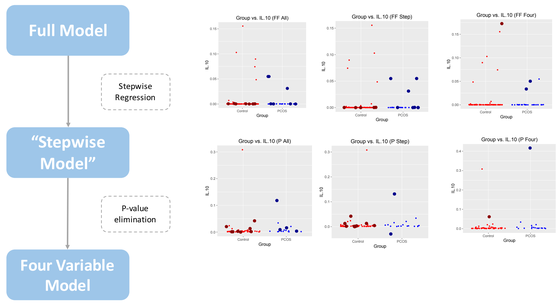
**Logistic Regression:** $\ln\left(\frac{\vec{p}}{1-\vec{p}}\right) = \beta_0 + \beta_1 X$

**Linear Kernel:** $K(\vec{x}_1, \vec{x}_2) = (\vec{x}_1 \cdot \vec{x}_2)$

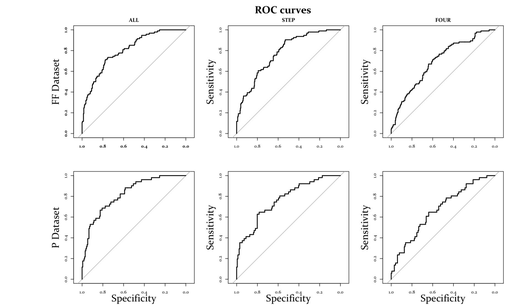
**SVM Optimization Problem:**

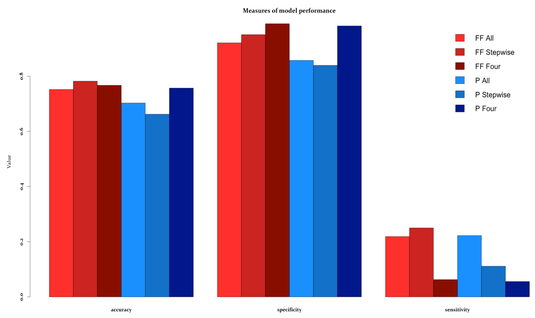
$$f(x) = sign(w \cdot x + b)$$

## Statistical Analysis



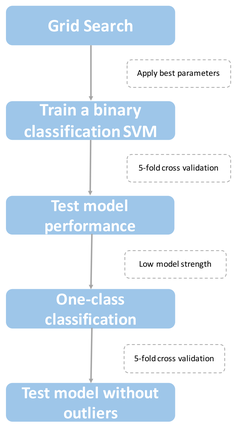

First, we put all 48 variables in a logistic regression model and predicted the probability of each sample being a PCOS or Control **(ALL).** Both datasets had comparably poor performance. Then we conducted stepwise regression, which uses the **AIC criteria** to eliminate the least significant variable **(STEP)**. Once this process was complete, we analyzed the performance of the model using the FF and P datasets once again. Finally, to get a sense of how the models would perform with a more clinically feasible number of variables, we removed variables based on the highest p-value until we were left with just four cytokines **(FOUR). ROC curves** highlighting the performance of all six models are displayed below. In addition, a visual of accuracy, specificity, and sensitivity are also displayed.



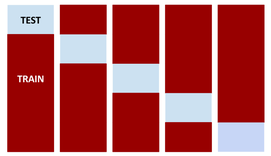

| Model | # Variables | Training Set | Testing Set | Specificity | Sensitivity | Accuracy |
|---|---|---|---|---|---|---|
| FF All | 48 | 75% of full FF | 25% of full FF | 0.9207921 | 0.21875 | 0.7518797 |
| FF Stepwise | 21 | 75% of full FF | 25% of full FF | 0.950495 | 0.25 | 0.7819549 |
| FF Four | 4 | 75% of full FF | 25% of full FF | 0.990099 | 0.0625 | 0.7669173 |
| P All | 48 | 75% of full P | 25% of full P | 0.8571429 | 0.2222222 | 0.7027027 |
| P Stepwise | 12 | 75% of full P | 25% of full P | 0.8392857 | 0.1111111 | 0.6621622 |
| P Four | 4 | 75% of full P | 25% of full P | 0.9821429 | 0.05555556 | 0.7567568 |

## Using Support Vector Machines

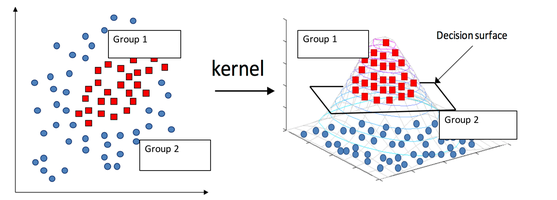


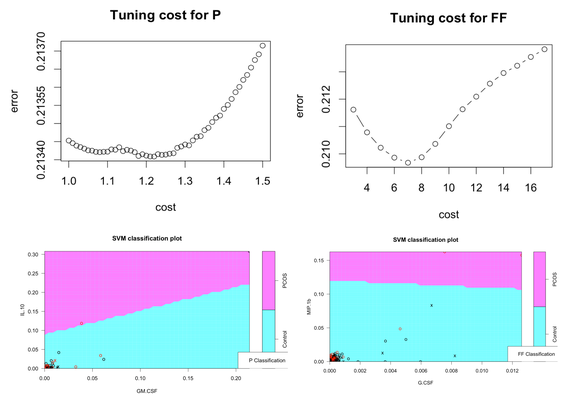

First we performed grid search in order to optimize parameters for our SVM model using a linear kernel. Then, we trained a binary classifier and tested its performance using 5-fold cross validation (results in the table below). The results led us to conduct one-class classification for outlier detection. Then we retested our model excluding the points we found to be outliers.

### 5-fold Cross Validation




### How SVM Works[1]

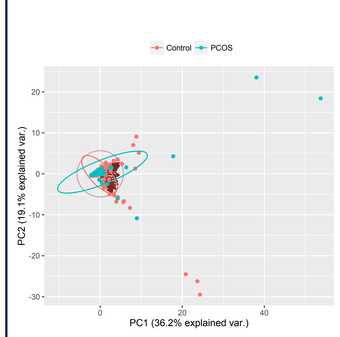



| Model | # Variables | Training Set | Testing Set | Specificity | Sensitivity | Accuracy |
|---|---|---|---|---|---|---|
| FF SVM | 7 | 75% of full FF | 25% of full FF | 1 | 0.0625 | 0.7744361 |
| | | 5-fold CV | 5-fold CV | 0.9876543 | 0.03846154 | 0.7570093 |
| | | 5-fold CV (no outliers) | 5-fold CV (no outliers) | 1 | 0 | 0.7583333 |
| P SVM | 3 | 75% of full FF | 25% of full FF | 0.9821429 | 0.05555556 | 0.7567568 |
| | | 5-fold CV | 5-fold CV | 1 | 0 | 0.7627119 |
| | | 5-fold CV (no outliers) | 5-fold CV (no outliers) | 1 | 0.06666667 | 0.7878788 |

## Other Analyses

### Principal Component Analysis (PCA):




PCA uses an orthogonal transformation to make a set of variables linearly independent variables called principal components. The above line graph represents how much variance in the data is accounted for by each principal component. The graph to the left represents the groupings based on the two most significant principal components.

### K-Means Clustering:

K-means clustering is a stochastic process that groups data points based on their distance from each other. Points are randomly assigned to a cluster, then cluster placement is optimized by finding the center of each cluster. Our clustering analysis resulted in the graph to the right, which is highly condensed because of the presence of so many unique outliers, which is consistent with one-class classification.

## Future Directions

The dataset described in this study is a small subset of a much larger collection of patient data. In the future we plan to incorporate more of these data to determine what more can be said about the classification of PCOS and prediction of fertility treatment success. Another question we might be interested in is if Follicular Fluids data could be better used to predict pregnancy results while blood plasma data could be better at predicting the presence of PCOS. Applying different methods of analysis (i.e. different machine learning classifiers) may have different strengths than our current models.

## References

1. http://www.med.nyu.edu/chibi/sites/default/files/chibi/Final.pdf
2. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3872139/
3. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
4. Latchman, David S. Gene Regulation- Fifth Edition. Taylor & Francis Group 2005. Print.
5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2785020/

## Acknowledgements

Thank you to the members of the Gunawardena lab including but not limited to John, Sieu, Dan, David, Deepesh, Mohan, Felix, Javi.
This work was supported by the Gwill York and Paul Maeder Research Award for Systems Biology and the FAS Center for Systems Biology. This work was supported by the National Science Foundation, Award id. 1462629.