Use of Machine Learning Algorithms to Predict Illness Severity and Diagnosis Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) From Patient Data

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364879">https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364879</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Use of Machine Learning Algorithms to Predict Illness Severity and Diagnosis
Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) From Patient Data

Rill Friedman

A Thesis in the Field of Mathematics and Computation
for the Degree of Master of Liberal Arts in Extension Studies

Harvard University
March 2020

Copyright 2020 Rill Friedman
Abstract

Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) is a condition that is characterized by a constellation of symptoms, including post-exertional malaise and disabling fatigue. Neither a blood test nor a single set of biomarkers exists for ME/CFS.

Dr. Jose G. Montoya, former Professor of Medicine, Infectious Diseases, and Geographic Medicine at the Stanford University Medical Center, provided us with a multidimensional data set consisting of samples from 192 ME/CFS cases and 392 healthy controls. He eliminated participants with missing data, yielding a final sample size of 186 ME/CFS cases and 388 healthy controls. Each participant completed the MFI-20, a 20-item quality of life questionnaire, on the day of blood sample collection. Dr. Montoya measured each participant’s serum cytokine levels using a 51-multiplex array.

We divided this data into training and test sets to develop a binary predictive classifier to identify whether a study participant has ME/CFS. We tested eight different binary classification models. We fine-tuned the top three performing baseline models. We then employed the best of these three models, an elastic net generalized linear model, binary classification algorithm to determine whether a study participant has ME/CFS. The model’s total accuracy is 75%, with an area under the return operator curve of 0.703, sensitivity of 0.439, and specificity of 0.93. This model is statistically significant at the $\alpha < 0.05$ level with a McNemar's Test p-Value of 0.001.
Next, we similarly evaluated eight regression prediction models for disease severity. As above, we tuned the best three of the eight baseline models. The top performer of these three models, evaluated on mean absolute error (MAE), was partial least squares regression. This two-component partial least squares model has a root mean square error (RMSE) of 25.33, a training coefficient of multiple determination $R^2$ of 0.062, and a training MAE of 22.49, test RMSE of 25.86, and a test MAE of 23.25.

We successfully created a statistically significant binary classification model based upon 51 provided cytokine values in serum samples from ME/CFS patients and healthy controls. From the same cytokine values, combined with the MFI-20 survey measure of illness severity, we derived a regression and prediction model for illness severity. These findings may contribute to the development of a blood test for ME/CFS as well as the continuing hunt for disease biomarkers.
Author’s Biographical Sketch

Rill Friedman is a versatile computational model builder. They currently focus on biostatistics, machine learning, bioinformatics, and artificial intelligence.

Unfortunately, Rill contracted infectious mononucleosis in the Fall of 1992, and never fully recovered. They present a textbook case of post-viral ME/CFS and hope that their current research facilitates quicker and more accurate diagnoses, as well as a better understanding of the biological mechanisms involved in ME/CFS. Better diagnosis technology may lead, in turn, to targeted treatment.

They have recently shifted their work from modeling aerospace systems to modeling biological ones. In their first career, they spent more than a decade as a contractor for the United States Navy. They tested rocket motors, designed, integrated, and tested both hardware and software, and developed guidance and navigation algorithms for ballistic missile interceptors. Yes, actually, they are a rocket scientist.

Rill completed their undergraduate education at Tufts University, receiving a Bachelor of Science degree in Mechanical Engineering, concentrating in thermodynamics and fluid mechanics.
Dedication

For Dr. Kenneth J. Friedman, Associate Professor of Physiology (retired), New Jersey Medical School, University of Medicine and Dentistry of New Jersey; also, the author’s father.

In 1993, this fantastic scientist began his career all over again, at a high professional cost to himself, when his child became ill with ME/CFS, a disease that was, at that time, nigh unto impossible to diagnose, rarely acknowledged by mainstream medicine, and very poorly understood.

Since then, he has become one of the premier names in ME/CFS research and advocacy, repeatedly serving on the national advisory board for ME/CFS advising the US Secretary of Health.

He continues, even in his retirement, to publish often, peer review articles, and edit journals.

Also, through severe relapses, career changes, and personal challenges, his support for his child has been unfailing.
Acknowledgments

We all stand on the shoulders of giants.

This project was built upon research by Dr. Jose Montoya, formerly of Stanford University, Dr. Gordon Broderick, Nova Southeastern University, and Dr. Nancy Klimas, also of Nova Southeastern University.

Gratitude, as well to George Box, who said both, “All models are wrong, but some are useful,” as well as, “It is inappropriate to be concerned about mice when tigers are abroad.” (Box, Journal of the American Statistical Society, 1976).
Table of Contents

Abstract ................................................................................................................................. 2

Author’s Biographical Sketch .......................................................................................... 4

Dedication ........................................................................................................................... 5

Acknowledgments ............................................................................................................. 6

List of Tables .................................................................................................................... 10

List of Figures .................................................................................................................. 11

Chapter 1: Project Description ....................................................................................... 13

Chapter 2: Background ................................................................................................... 15

  2.1 What is ME/CFS? ...................................................................................................... 15

  2.2 Hunting for ME/CFS Biomarkers .......................................................................... 20

  2.3 Machine Learning for Diagnosis and Classification ............................................. 23

  2.4 Workflow ................................................................................................................. 25

  2.5 Study Methodology ............................................................................................... 27

  2.6 Data Description ..................................................................................................... 28

  2.7 R Packages ............................................................................................................. 31

    2.7.1 Caret .................................................................................................................. 31

    2.7.2 Glmnet .............................................................................................................. 31

    2.7.3 Ggplot2 ............................................................................................................ 31

Chapter 3: Algorithms ..................................................................................................... 33
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Classification</td>
<td>33</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Elastic Net</td>
<td>33</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Random Forest (RF)</td>
<td>34</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Support Vector Machine (Polynomial Kernel)</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>Regression</td>
<td>35</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Partial Least Squares</td>
<td>35</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Bagged Trees (Regression)</td>
<td>36</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Elastic Net Regression</td>
<td>37</td>
</tr>
<tr>
<td>3.3</td>
<td>Feature Extraction</td>
<td>38</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Recursive Feature Elimination (RFE)</td>
<td>38</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Stepwise selection</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Chapter 4: Implementation and Analysis</td>
<td>40</td>
</tr>
<tr>
<td>4.1</td>
<td>Data Wrangling</td>
<td>40</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Data Cleaning</td>
<td>40</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Outlier removal</td>
<td>40</td>
</tr>
<tr>
<td>4.2</td>
<td>Data Visualization</td>
<td>42</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Correlation</td>
<td>42</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Principal Component Analysis</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Chapter 5: Model Development</td>
<td>48</td>
</tr>
<tr>
<td>5.1</td>
<td>Baseline binary classification models</td>
<td>48</td>
</tr>
<tr>
<td>5.2</td>
<td>Classification Model Tuning</td>
<td>49</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Support Vector Machine (Polynomial Kernel)</td>
<td>49</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Random Forest</td>
<td>52</td>
</tr>
</tbody>
</table>
5.2.3 Generalized Linear Model (Elastic Net) ........................................... 55
5.3: Final Binary Classification Model Selection ........................................ 59
5.4 Baseline Regression Models ..................................................................... 61
5.5 Regression Model Tuning ........................................................................ 63
   5.5.1 Partial Least Squares ........................................................................... 63
   5.5.2 Bagged Tree Regression ...................................................................... 64
   5.5.3 Generalized Linear Model Elastic Net Regression .............................. 65
5.6 Best Regression Model Selection .............................................................. 66

Chapter 6: Discussion .................................................................................. 68
   6.1 Experimental Design ............................................................................ 68
   6.2 Analysis Summary ................................................................................ 69
   6.3 Future Directions in Research ............................................................... 70

Chapter 7: References ................................................................................. 72
   Works Cited ................................................................................................ 72
   Works Consulted ....................................................................................... 77
List of Tables

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Cytokine Correlation with Illness Severity ........................................ 44</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2:</td>
<td>Baseline Classification Model Accuracy ......................................................................... 48</td>
</tr>
<tr>
<td>Table 3:</td>
<td>Tuned Model Performance ............................................................................................. 59</td>
</tr>
<tr>
<td>Table 4:</td>
<td>Baseline Regression Model Metrics ............................................................................... 61</td>
</tr>
<tr>
<td>Table 5</td>
<td>Bagged Tree Regression Best Variable Subset .................................................................. 64</td>
</tr>
<tr>
<td>Table 6:</td>
<td>GLM Elastic Net Regression Best Variable Subset .......................................................... 65</td>
</tr>
<tr>
<td>Table 7:</td>
<td>Tuned Regression Models Mean Absolute Error ................................................................ 66</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1: Institute of Medicine’s ME/CFS Diagnostic Protocol............................... 16
Figure 2a: Workflow ................................................................................................ 25
Figure 2b: Workflow continued............................................................................... 26
Figure 3: Illness Severity and Patient Class......................................................... 41
Figure 4: Participant Illness Severity and Class After Outlier Removal............... 41
Figure 5: Cytokine/Severity Pearson Coefficient Correlation Matrix..................... 43
Figure 7: PCA Variance Explained by Principal Components.............................. 45
Figure 8: Participants (All Classes) Plotted on PC1/PC2 Axes.............................. 46
Figure 9: Cytokine Contributions to PC 1 .............................................................. 46
Figure 10: Cytokine Contributions to PC 2 .............................................................. 47
Figure 11: Boxplot of Baseline Classification Model Accuracy ......................... 49
Figure 12: SVM Poly Cost Optimization ............................................................... 50
Figure 13: SVM Poly Kernel Confusion Matrix.................................................... 51
Figure 14: SVM Polynomial Kernel ROC (TPR/FPR) ......................................... 51
Figure 15: SVM Polynomial Kernel ROC (Sens/Spec) .......................................... 52
Figure 16: RF Model Accuracy vs. Number of Variables Available for Splitting.... 53
Figure 17: Optimized Random Forest Model Confusion Matrix........................... 54
Figure 18 Random Forest ROC (TPR/FPR).......................................................... 54
Figure 19: Random Forest ROC (Sensitivity/Specificity)........................................ 55
Figure 20: Optimization of Mixing of $\alpha$ and $\lambda$ Coefficients ....................................... 56
Figure 21: Best Elastic Net Model Confusion Matrix ...................................................... 57
Figure 22: Elastic Net Model ROC (TPR/FPR) ............................................................. 57
Figure 23: Elastic Net Model ROC (Sensitivity/Specificity) ........................................ 58
Figure 24: Tuned Model Performance ........................................................................... 60
Figure 25: Baseline Regression Models Mean Absolute Error ..................................... 62
Figure 26: PLS Regression Model Optimization of Number of Components .......... 64
Figure 27: Elastic Net Regression Parameter Optimization ......................................... 66
Figure 28: Tuned Regression Model MAE .................................................................... 67
Chapter 1: Project Description

ME/CFS is a devastating illness comprised of severe fatigue and a constellation of other symptoms. Over 800,000 adults in the US are affected. The estimated cost to society annually is $2 billion. (Jason et al., 2015) No single diagnostic test exists for ME/CFS, making the disease difficult to diagnose. The current diagnostic protocol appears below (Figure 1).

The etiology of ME/CFS also remains unknown, and many causes are currently under investigation studied. The Centers for Disease Control and Prevention (CDC) suggests in their paper on etiology and pathophysiology of ME/CFS that causes may include: infection, trauma, genetics, and environmental factors. (“Possible Causes, Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS), CDC,” 2017). A subset of ME/CFS patients develops the disease after infection with Epstein Barr Virus (infectious mononucleosis), Coxiella Burnetti (the causative agent of Q fever), and Ross River Virus. Other infectious agents, including mycoplasma, human herpes viruses, and retroviruses, are also being studied.

Another set of ME/CFS patients presents after significant physical or emotional trauma. Examples include motor vehicle accidents, stressful life events, immobilization, and surgery. Clusters of ME/CFS patients have been identified within families, thus suggesting a genetic predisposition toward the disease. Environmental factors, such as mold, are also suspected in some cases. (“Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS),” CDC, 2018)

By using serum samples from both ME/CFS patients and healthy controls to create both a binary classification and prediction model and a disease severity regression and
prediction model, this project contributes to the refinement of the ME/CFS diagnostic process. Identification of distinct biomarkers in ME/CFS patients is crucial to the development of a blood test for the disease.
Chapter 2: Background

2.1 What is ME/CFS?

The Solve ME/CFS Initiative, in the "What is CFS?" section of their website, describes ME/CFS as "a complex and debilitating multi-system, chronic disease with a serious impact on one's quality of life." They identify a constellation of symptoms, including the hallmark of ME/CFS, extreme exhaustion made worse by either physical or mental effort. This phenomenon, unique to ME/CFS, is known as post-exertional malaise. Also included are non-restorative sleep, brain fog, and cognitive impairment, joint pain, inflamed lymph nodes, persistent sore throat, severe headache, other neurological abnormalities, sensitivity to light, sound, medications, chemicals, smell, and occasionally complete organ system shutdown. Symptoms may also include sensitivity to light, sound, odors, chemicals, foods, and medications (“About the Disease,” Solve ME/CFS Initiative, retrieved December 2018.)

In 2015, the Institute of Medicine published the following diagnostic algorithm, meaning that the diagnosis of ME/CFS no longer happens by exclusion.
Figure 1: Institute of Medicine’s ME/CFS Diagnostic Protocol
Diagnosis depends upon thorough physical exam, history, exclusion of other illnesses, and confirmatory testing. Although no single biomarker test exists, there are test results consistent with a diagnosis of ME/CFS.

The American Myalgic Encephalomyelitis and Chronic Fatigue Syndrome Society states that upon exam, a physician may note a variety of observable phenomena. They may see low blood pressure (with orthostatic hypotension confirmed via tilt table test). Also frequent is abnormal body temperature, low (97°F), high (100°F), or, more often, an excessive diurnal variation of temperature. Tachycardia, which is hard to detect during an office visit, may be found via Holter monitor. Patients’ throats are often irritated, and lymph nodes, particularly on the left side, are often tender. Dr. Cheney notes that his Caucasian patients often possess a ghostly pallor. They may also have a stiff, slow, gait, nystagmus, photophobia, or hyperreflexia. ("About the Disease," Solve ME/CFS Initiative, retrieved September 2019)

In addition to the above physically observable phenomenon, routine lab tests may show either decreased or increased white blood cells. Red blood cell membranes may be abnormal, as well. Patients often have low concentrations of zinc and magnesium, low uric acid concentration, slightly elevated total cholesterol, and decreased potassium and sodium. Their sedimentation rates are often low (<5 mm/hr). Yet, sometimes brief periods of sedimentation rate elevation (> 20 mm/hr) occur. Urinalysis results may be alkaline, with mucus, blood, or both, but without bacterial infection. Antinuclear Antibodies (ANA) may be positive in a speckled pattern. Liver function tests may show mildly elevated aspartate aminotransferase (AST) serum glutamic-oxaloacetic transaminase, also known
as (SGOT), and alanine aminotransferase (ALT) formerly known as serum glutamic-pyruvic transaminase, (SGPT).

Some specialized tests may indicate a diagnosis of ME/CFS. In the 2 Day Cardiopulmonary Exercise Test (CPET), ME/CFS patients show decreased cortisol levels after exercise, decreased cerebral blood flow after exercise, inefficient glucose utilization, and an erratic breathing pattern. When measuring maximal oxygen uptake (VO2 max) during exercise, ME/CFS patients score significantly lower than controls.

Note that exercise testing excludes both moderately and severely ill patients, as it assumes that the patient can get to the test site, stand up long enough to do the test, and exercise both within and beyond their aerobic threshold.

ME/CFS patients may test positive for viral reactivations, including Cytomegalovirus (CMV), Epstein-Barr Virus (EBV), Human Herpes Virus 6 (HHV6), Human Herpes Virus 7 (HHV7), and Coxsackie Virus. They simultaneously tend to test negative for the Hepatitis A or B viruses.

There are a variety of immune system abnormalities common among patients with ME/CFS. Patients tend to display low natural killer cell counts. Interferon-alpha, tumor necrosis factor, in addition to interleukins 1 and 2, may be elevated. T cells may be activated, and the T-Lymphocyte Helper/Suppressor Profile (T4/T8) ratio may be altered. T cell suppressor cell (T8) count trends low, whole T cell count fluctuates between low and high. B cell count may also be either low or high. Antinuclear antibodies, immunoglobulin deficiency, and sometimes antithyroid antibodies may also be present.
Brain scans of ME/CFS patients may be abnormal. Single-photon emission computed tomography (SPECT) brain scan may show hypoperfusion in either the left or right temporal lobe, particularly Wernicke’s area, after exercise.

Magnetic resonance imaging (MRI) may show unexplained bright areas. Magnetic resonance spectroscopy (MRS) may show abnormally high lactic acid spikes, indicating mitochondrial dysfunction, near and around the hippocampus. Heart function, in addition to the tachycardia mentioned above, may display abnormal impedance cardiography, repetitive oscillating T wave inversions, or T wave flats during Holter monitoring. (“How Is ME/CFS Diagnosed?” American ME and CFS Society, retrieved December 2018.)

The status of the hunt for biomarkers and immunological patterns, unique to many ME/CFS patients, is covered in the "State of the Science, ME/CFS" section of this paper. The Solve ME/CFS Initiative estimates the prevalence of the disease at “between 836,000 and 2.5 million Americans” and “17 – 24 million people with ME/CFS worldwide.” There is no FDA approved treatment for ME/CFS, and patients are currently treated symptomatically.
2.2 Hunting for ME/CFS Biomarkers

The hunt for biomarkers for ME/CFS is ongoing. In 2011, Benu and Klimas et al. published an “Immunological abnormalities as potential biomarkers in ME/CFS.” Their research found several possible biomarkers, demonstrating significant immune dysregulation. They found “significant increases in IL-10, IFN-γ, TNF-α, CD4+CD25+ T cells, FoxP3 and VPACR2 expression” as well as significant decreases in “Cytotoxic activity of NK and CD8+T cells and NK phenotypes, in particular, the CD56 bright NK cells.” They also found that “granzyme A and granzyme K expression were reduced while expression levels of perforin were significantly increased in the CFS/ME population relative to the control population.” (Benu, Klimas et al., 2011)

A 2015 study by Landi, et al., examined 100 ME/CFS long duration of illness patients, and 79 matched controls. They found highly significant reductions in circulating inflammatory cytokines IL 16, IL 7, and VEGF-A. They also pointed out that measurement of the same analytes in patients with both chronic infectious and autoimmune liver diseases, situations with similar fatigue-based symptomology, failed to show the same decrease. They thus concluded that the ME/CFS cytokine profile might be unique. They called for further study of diseases with overlapping morphology. (Landi, et al., 2015)

Another 2015 paper, by Peterson, found significantly decreased levels of cytokine IL-10 in the cerebrospinal fluid of ME/CFS patients. Interestingly, they did not find reduced levels of IL 7. IL 16 and VEGF-A were absent from this study. (Peterson, et al., 2015)

In 2016, Russell, Broderick, et al. were able to build a diagnostic classification model for ME/CFS using a generalized linear model. They discovered that the expression
of discriminatory cytokines shifts with both age and duration of illness. They focused their study on IL-1α, 6, and 8. "Setting these three markers as a triple screen and adjusting their contribution according to illness duration sub-groups produced ME/CFS classification accuracies of 75–88 % using a 64 plex cytokine assay.

Their work demonstrated that the contribution of IL-1α, higher in recently ill adolescent ME/CFS subjects, was progressively less critical with illness duration. While high levels of IL-8 screened positive for ME/CFS in the recently afflicted, the opposite was true for subjects ill for more than two years. Similarly, while lower levels of IL-6 suggested early ME/CFS, the reverse was true in participants over 18 years of age, ill for more than two years." (Russell et al., 2016)

Hornig et al. studied the levels of inflammatory cytokines in the cerebrospinal fluid (CSF) of patients with ME/CFS, Multiple Sclerosis (MS), and normal controls. This study divided the ME/CFS patients into two categories, typical (long duration, group C) and atypical (short duration, group A) cases. They utilized both Random Forest (RF) and logistic regression to find statistically significant differences between long and short duration ME/CFS patients. Short duration cases have lower levels of IL7, IL1β, IL5, IL7, IL13, IL17A, IFNα2, IFNγ, TNFα, TRAIL (TNFSF10), CCL2, CCL7, CXCL5, CXCL9, CSF3 (GCSF), βNGF, resistin and serpin E1 (PAI1). The only cytokine that has a higher level in short duration cases was FGFb. Hornig, and team, also used network association diagrams to reveal different interrelationships among CFS cytokines between ME/CSF patient subgroupings. (Hornig, et al. Immune Network Analysis of Cerebrospinal Fluid in ME/CFS, 2017)
Montoya et al. correlated a specific cytokine signature with illness severity. His group found that 17 specific cytokines displayed a strongly correlated linear upward trend with disease severity. (Montoya et al., 2017)

Other biostatistical research on ME/CFS uses plasma metabolites instead of cytokines and other inflammation markers to discriminate between ME/CFS patients and healthy controls. This work includes a July 2018 publication, “Insights into myalgic encephalomyelitis/chronic fatigue syndrome phenotypes through comprehensive metabolomics.” This paper summarizes the analysis of 562 plasma metabolites. The research described uses ten random forests ranked metabolites fitted as predictors in the predictive multivariate logistic regression models." The predictive model was able to distinguish ME/CFS patients from controls with “high accuracy.” (Nagy-Szakal et al., 2018)

Recent research by Xu et al. has examined single-cell Raman spectra (SCRS) profiles. The experimental results demonstrate that “Raman bands associated with phenylalanine in $\rho^0$ cells and CFS patient peripheral blood mononuclear cells (PBMCs) were significantly higher than those of the wild-type model and healthy controls. As similar changes occurred in the $\rho^0$ cell model with a known deficiency in the mitochondrial respiratory chain as well as in CFS patients, our results suggest that the increase in cellular phenylalanine may be related to mitochondrial/energetic dysfunction in both systems. A non-specified machine learning binary classification model was able to differentiate CFS patients from controls with 98% accuracy based on Raman spectra. (Xu et al., 2018)

Stanford University has an ongoing effort led by Ron Davis, Ph.D., and Xenzhong Xiao, Ph.D., entitled the "Severely Ill Big Data Study." There are 1000 test points per
patient, including genome, gene expression, metabolomics, microbiome, cytokines, cell-free DNA sequencing and quantitation, clinical test results, and cytokines. This study is the largest ME/CFS patient data set to date. So far, differences in clinical test results, cytokines, and microbiomes separate patients and controls. Also, of note is the finding that quality of life, as demonstrated by the short item survey (SF-36) score, is worse for ME/CFS patients than several other major diseases. It also correlates least with depression and mental illness. (RAND, 36 Item Short Survey, OMF, 2018)

2.3 Machine Learning for Diagnosis and Classification

The 2013 paper by Kumari and Chitra detailing the results of their effort to build a machine learning classifier for the diagnosis of diabetes inspired this effort. They were able to obtain an accuracy of 75% using a Support Vector Machine model with a Radial Basis Function (RBF) kernel. The team analyzed a high dimensional data set. (Kumari and Chitra, 2013)

Mexican scientists, Eddy Sanchez-DelaCruz and Pilar Pozos-Parra, have used a multiclass artificial neural net (ANN), multilayer perceptron, one vs. all model, to identify patients with the neurodegenerative diseases Parkinson's Disease (PD), Huntington's (HD), and Amyotrophic lateral sclerosis (ALS) with 84% accuracy. Their work builds on the publicly accessible Gait ND-DB data set. It contains 13 measured variables describing each patient’s walking gate. There were 16 healthy controls, 15 patients with Parkinson’s Disease, 20 with Huntington’s, and 13 with ALS. (Sanchez-Delacruz and Pozos-Parra)

In 2017, Meherwar Fatima and Maruf Pasha published a Survey of Machine Learning Algorithms for Disease Diagnostic. They found a wide variety of highly accurate models evolving to diagnose heart disease, breast cancer, diabetes, Dengue disease, and
hepatitis. The algorithms most commonly used were Naive Bayes and Support Vector Machines (SVM). Also included were Artificial Neural Net (ANN), and Random Forest (RF). (Fatima and Pasha, 2017)

A 2018 paper, "Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables," published in the Lancet, supported these efforts. That work demonstrates the usage of a k-Means algorithm and hierarchical clustering to derive a robust set of 5 clusters from 3 independent data sets. The stated goal of the ongoing research is the stratification of patients into subtypes for targeted treatment. (Ahlqvist, Storm, et al., 2018)
2.4 Workflow

Figure 2a: Workflow

_Detailed workflow for data analysis_
Figure 2b: Workflow continued

*Figure 2b: Detailed workflow for model building and prediction.*
Note: On abbreviations, PPV is Positive Predictive Value, NPV is Negative Predictive Value, PC is Principal Component, MAE is Mean Accuracy Error, RMSE is Root Mean Square Error.

2.5 Study Methodology

Our work builds upon Dr. Montoya’s data set. His study methodology is excerpted below.

"Study Design. An age- and sex-matched case-control cross-sectional study was conducted at Stanford University in 2009 to investigate the role of immune responses, genetic predisposition, and infection in ME/CFS. A total of 192 ME/CFS cases and 392 healthy controls are included in this analysis. However, the final sample size was 186 ME/CFS cases and 388 healthy controls because of missing race information in six ME/CFS cases and four healthy controls. In addition to the serum analyzed in this report, other peripheral blood specimens were collected from these individuals and remain available for ongoing and future immune, genetic, and pathogen-discovery studies.

Each case participant was matched to two control participants by sex and age (±6 mo). To be included in the study, participants had to be 14 y of age or older, reside in Northern California, and provide written informed consent and Health Insurance Portability and Accountability Act of 1996 authorization as required by the Stanford University Institutional Review Board (protocol numbers 18068 and 18155). Participants were classified as cases if they met the 1994 CDC CFS case definition (3). Of note, symptoms such as unrefreshing sleep, postexertional malaise, and impaired memory (also referred to by patients as "brain fog") were present in 96.9, 96.9, and 95.8% of ME/CFS patients, respectively (Table 1)

Controls in the study were eligible if they did not have a history of fatigue and did not meet the ME/CFS case definition. Exclusion criteria for both groups included active or uncontrolled morbidities that would have interfered with the patient's ability to participate in the study, particularly conditions or medications causing immunosuppression or immunodeficiency.

Participants were recruited from March 2, 2010, to September 1, 2011, and their peripheral blood was drawn between 8:30 AM and 3:30 PM on the day of enrollment. Blood specimens including serum, plasma, whole-blood DNA and whole-blood RNA (collected in PAXgene tubes), and PBMCs were obtained and processed on the day of enrollment by the
Participants were recruited from March 2, 2010, to September 1, 2011, and their peripheral blood was drawn between 8:30 AM and 3:30 PM on the day of enrollment.

The MFI-20, a 20-item questionnaire (51), was administered to each participant on the day of blood sample collection. A higher score indicates greater severity. This instrument has been validated in the ME/CFS population (58). Cytokine Assay. Cytokines were measured for each participant in serum using a 51-multiplex array on the Luminex 200 IS system (Affymetrix) performed at the Stanford HIMC. The manufacturer’s protocol was followed, with variations as described by Brodin, et al. (59). A total of 19 plates were used. Each participant’s sample was entered in two replicate wells, and matched sets of ME/CFS cases and healthy controls were always mixed in all plates to reduce confounding case status with plate artifacts. Results were accepted as final (569 samples) if more than 95% of data had a coefficient of variation (CV)” (Montoya et al., 2017)

2.6 Data Description

Dr. Montoya provided his data in a preprocessed MS Excel spreadsheet with an accompanying MS Excel data dictionary.

The data entry for each study participant is described below in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Variables in Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable Name</strong></td>
</tr>
<tr>
<td>id</td>
</tr>
<tr>
<td>match_set</td>
</tr>
<tr>
<td>case_control_status</td>
</tr>
<tr>
<td>sex</td>
</tr>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
<tr>
<td>l=male</td>
</tr>
<tr>
<td>total</td>
</tr>
<tr>
<td>mfi20cat</td>
</tr>
<tr>
<td>Race</td>
</tr>
<tr>
<td>Full_Cytokine_Name</td>
</tr>
<tr>
<td>Preprocessed_mfi</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>CHEX4_REMOVED_mfi</td>
</tr>
</tbody>
</table>

*Table 1 defines variables contained within the data set*
Note. The Multidimensional Fatigue Inventory (MFI-20) is a survey tool designed to measure fatigue severity. The MFI-20 consists of five subscales: general fatigue, physical fatigue, mental fatigue, reduced activity, and reduced motivation. Each subscale includes four items scored on a five-point scale. Scores on each subscale range from 4 to 20, with higher scores indicating greater fatigue. (Lin et al., 2009)

Note. • Mendel and Mendel discuss analytical compensation for nonspecific binding in their article, "Non-specific Binding, the Problem, and a Solution." (Mendel & Mendel, 1985)
2.7 R Packages

2.7.1 Caret

Classification and REgression Training

author: Max Kuhn with Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. (Caret Package Documentation, Retrieved October 2019)

Caret contains functions that streamline model training and testing for classification and regression problems. The package utilizes a variety of other R packages but does not load them on startup. It calls them as needed. Caret contains a sleek interface for training, data partitioning, model tuning, and prediction.

2.7.2 Glmnet

authors: Jerome Friedman, Trevor Hastie, Rob Tibshirani, and Noah Simon

Glmnet fits a generalized linear model via penalized maximum likelihood. Coefficients are computed, for lasso or elastic net penalty, at a grid of values for the regularization parameter $\lambda$. (Hastie and Qian, Glmnet Vignette, Retrieved October 2019)

2.7.3 Ggplot2

“Ggplot2 is a system for declaratively creating graphics, based on “The Grammar of Graphics. The user provides the data and tells it how to map the variables to aesthetics, what graphical primitives to use, and it takes care of the rest.”
authors: Hadley Wickham and Winston Chang

(Wickham and Chang, Ggplot2, retrieved October 2019)
Chapter 3: Algorithms

There are two separate but related tasks in this analysis: classification and severity prediction. The binary classification task partitions data into one of two predefined classes (positive and negative).

3.1 Classification

3.1.1 Elastic Net

An elastic net is a form of generalized logistic model is a form of penalized logistic regression. The algorithm imposes a penalty to the logistic model for having too many variables. This penalty results in shrinking the coefficients of the less contributive variables toward zero, thus eliminating the less useful variables. This variable space dimensional reduction is also known as regularization.

Elastic net regression is a combination and optimization of ridge and lasso regression. Using this form of logistic regression for prediction, we estimate the coefficients $\alpha$ and $\lambda$ using maximum likelihood.

Using a grid search, we find the optimum combination of coefficients and apply these to the original logistic function. These coefficient estimates, gained from the maximum likelihood function are used with the test data to predict both patient classification (STDA) Classification Methods Essentials: Penalized Logistic Regression (Ridge, Lasso, and Elastic Net), (Zou and Hastie, 2005)
3.1.2 Random Forest (RF)

“Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.” (Breiman, UC Berkeley, 2001)

This method utilizes the wisdom of the crowd over the individual, hence the forest analogy. By using many trees, rather than one, the generalization error of this method converges, asymptotically, to a limit as the number of trees increases.

The classification power of a random forest depends on the decision strength of individual trees, and how well they correlate with each other. The model monitors error, strength, variable importance, and correlation internally. Selection of the optimal number of features available for node splitting (mtry). is based upon these factors. (Breiman, 1996)

3.1.3 Support Vector Machine (Polynomial Kernel)

SVMs are maximum margin classifiers. They maximize the margin around the separating hyperplane, the distance from the data point to the separator or boundary. (Gordon, 2004)

The decision function is fully specified by a small subset of training samples, the support vectors. The points closest to the decision surface form the support vectors, have the greatest influence on the hyperplane, and are also the most difficult to classify.

If the data set is linearly separable, in two dimensions, it can be separated by a line. These linear cases use an “Inner Product Kernel” of the form: \( K(x_1x_i) \) I=1, 2,... n. In higher dimensional linearly separable cases, we use hyperplanes.
If the data is not linearly separable, we can sometimes transform it into a linearly separable form by mapping the data to a higher-dimensional space. Transformation into polar coordinates is an excellent example of this. This type of transformation is known as the "kernel trick."

The polynomial kernel can be expressed, \( k(x, z) = (x \cdot z)^d \). The exponent \( d \) is provided a priori by the user. ("An Idiot's Guide to Support Vector Machines," Berwick, retrieved September 2019)

3.2 Regression

3.2.1 Partial Least Squares

Partial Least Squares Regression (PLSR) is an extension of multiple linear regression. Multiple Linear Regression (MLR) is limited by the imposition of the restriction that factors underlying the X and Y variables are extracted from the \( X'X \) and \( Y'Y \), respectively, and not from the cross product matrices that involve both X and Y. MLR is also constrained so that the number of prediction functions cannot exceed the minimum number of X and Y variables.

In contrast, PLSR lifts the first restriction and extracts factors form the \( Y'XX'Y \) matrix. The number of prediction functions so extracted typically exceeds the maximum number of X and Y variables. The flexibility of PLSR allows for its use in situations where traditional multivariable methods are limited. It is useful when there are fewer observations than predictor variables, as well as for selecting variables and identifying outliers before
linear regression. (StatSoft Textbook, Partial Linear Regression, retrieved September 2019)

3.2.2 Bagged Trees (Regression)

Tree-structured regression is an approach to the regression problem that is not based upon the assumption of normality. Tree-based models are ideal for situations where the relationship between the predictors and the outcome are nonlinear, or where features interact with each other. Tree models split, or subset, the original data set multiple times based upon cutoff values in the features. Each instance can belong to only one subset. Intermediate subsets are called internal or split nodes. Final subsets are terminal or leaf nodes. Predicting the outcome of a leaf node is performed by averaging the outcome of the training data in the node. (Interpretable Machine Learning, 4.4 Decision Tree). These intermediate node values differentiate trees from discriminant analysis and ordinary least squares. Tree-structured predictors are far simpler functions than kernel-based methods and nearest neighbors.

Bootstrap Aggregation (Bagging) is a general technique for improving prediction rules. It is an example of what Breiman calls a “perturb and combine” method. (Breiman, 1998). Here, a regression (or classification) method is applied to various perturbations of the original data set. The results are combined to create a single model. Like Random Forest, this method employs the wisdom of the masses over the knowledge of the individual. In this case, many bootstrap samples are drawn from the original data. A prediction method, in this case, a regression tree, is applied to each bootstrap sample. We then average the results. (Sutton, 2005)
3.2.3 Elastic Net Regression

An Elastic Net is a form of generalized logistic model; it is a form of penalized logistic regression. The algorithm imposes a penalty to the logistic model for having too many variables. This penalty results in shrinking the coefficients of the less contributive variables toward zero, thus eliminating the less useful variables. This variable space dimensional reduction is also known as regularization.

Elastic net regression is a combination and optimization of ridge and lasso regression. Using this form of logistic regression for prediction, we estimate the coefficients $\alpha$ and $\lambda$ using maximum likelihood. Using a grid search, we find the optimum combination of coefficients and apply these to the original logistic function. These coefficient estimates, gained from the maximum likelihood function are used with the test data to predict both patient classification (STDA Classification Methods Essentials: Penalized Logistic Regression (Ridge, Lasso, and Elastic Net), retrieved September 2019), (Zou and Hastie, 2005)
3.3 Feature Extraction

3.3.1 Recursive Feature Elimination (RFE)

Recursive Feature Elimination is a backward selection process that begins by building a model based upon the entire set of predictors and computing an importance score for each. The algorithm then removes the “least important” predictor(s) and rebuilds the model. In practice, the user/analyst has the option to set the number of predictor subsets, and the number of predictors per subset as tuning parameters. This method is of limited utility when predictor variables are highly correlated. When multicollinearity is present, the method may include all similar variables. (Johnson and Kuhn, 11.3 Recursive Feature Elimination (RFE), retrieved September 2019)

3.3.2 Stepwise selection

Stepwise Feature Selection is a forward stepwise regression approach developed for linear models. Predictor variables enter the model one at a time if they meet the entry criterion. Typically, the model holds a threshold where a variable may only enter if it has a p-value of < 0.15.

The algorithm creates p linear regression models, one for each predictor feature. We then rank the importance of said features by p-value. The best variable, with the lowest p-value, is added to the model if it meets the threshold criterion. After running the conglomerate model, we select and a metric such as $R^2$, Aiken Information Criterion(AIC), Root Mean Square Error (RMSE), or Mean Accuracy Error (MAE) to examine.
The next iteration builds p-1 linear regression models. The model stops when either of two conditions occurs: the metric stops improving, or no more variables meet the entry criterion. Stepwise feature selection is criticized because the method tends to over select features, particularly highly correlated ones. Over selection results in model overfitting. (Johnson and Kuhn, 11.4 Stepwise Selection, retrieved September 2019)
4.1 Data Wrangling

We imported the data into R and converted it from “long” to “short” form by creating one row per participant, including total severity score, participant class, and the 51 cytokine values.

4.1.1 Data Cleaning

The next task was data cleaning. We checked for missing values. There were none. We next plotted disease severity vs. class, finding that there were participants with high illness severity scores in the control class data.

4.1.2 Outlier removal

Figure 2 below illustrates the fact that there are moderately ill and severely ill study participants in the control class. These participants will confound any analysis of severity, particularly regression. We removed these participants using R’s built-in outlier elimination function. Figure 3 shows the participant distribution by class after outlier removal.
Figure 3: Illness Severity and Patient Class

Note: See the left-hand column above, where moderately and severely ill patients are in the control class (0). Outliers display as the data points that lie above the whiskers. Sick participants in the "healthy control," particularly severely ill ones, confound all attempts at regression. The severity of illness in the control group does not affect the binary classification problem.

Figure 4: Participant Illness Severity and Class After Outlier Removal
Note: In Figure 4, outlying data points have been removed from the control group (class 0) only. Without moderately and severely ill participants in the control category, regression analysis was possible.

We were able to ascertain, via discussion, with Dr. Montoya’s lab’s statistician, that the 51 cytokine values for each participant had already been log-transformed, scaled, and centered. The illness severity scores, also for each participant, ranged from 4 to 100.

4.2 Data Visualization:

4.2.1 Correlation

The first thing we examined was the structure of the data, looking for any visually apparent trends or patterns. We calculated a correlation matrix for all 51 cytokines and illness severity, looking for correlations between illness severity and cytokines. As shown in the correlation matrix plot below, we were able to confirm some of the conclusions made by Montoya et al., in their 2017 PNAS paper. We found that 10 of the cytokines have statistically significant linear correlations with severity.
Figure 5: Cytokine/Severity Pearson Coefficient Correlation Matrix

The correlation matrix shows multicollinearity between cytokines as well as a statistically significant correlation between 10 cytokines and illness severity.
Cytokine Correlation with Illness Severity

Table 1  Cytokine Correlation with Illness Severity

<table>
<thead>
<tr>
<th>Cytokine</th>
<th>Pearson Coeff</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCL5</td>
<td>0.1042803</td>
<td>0.0138052</td>
</tr>
<tr>
<td>CD40L</td>
<td>-0.0958637</td>
<td>0.0236607</td>
</tr>
<tr>
<td>CXCL5</td>
<td>-0.0862265</td>
<td>0.0419276</td>
</tr>
<tr>
<td>IL13</td>
<td>0.1150865</td>
<td>0.0065464</td>
</tr>
<tr>
<td>IL2</td>
<td>0.1114022</td>
<td>0.0085013</td>
</tr>
<tr>
<td>IL5</td>
<td>0.0835525</td>
<td>0.0487322</td>
</tr>
<tr>
<td>NGF</td>
<td>0.1125373</td>
<td>0.0078499</td>
</tr>
<tr>
<td>PAI.1</td>
<td>0.0937712</td>
<td>0.0268986</td>
</tr>
<tr>
<td>RESISTIN</td>
<td>-0.1285903</td>
<td>0.0023606</td>
</tr>
<tr>
<td>TGF:beta</td>
<td>0.1526104</td>
<td>0.0003005</td>
</tr>
</tbody>
</table>

Ten cytokines show a statistically significant correlation with Illness Severity

We derived these correlation numbers by comparison between each cytokine and severity. Multiple comparisons lead to an increased probability of false inference, increased false-positive rate, for which there are several methods of compensation. The most common of these is the Bonferroni correction wherein we divide $P=0.05$, our desired statistical threshold for $\alpha$, by the number of tests (51) to get the Bonferroni critical value 0.00098 or approximately $P<0.001$. Placing the significance threshold at $P < 0.001$ is an extremely conservative requirement. When using this criterion for statistical significance, only one of the cytokines had a statistically significant correlation.

McDonald states in his Handbook of Biological Statistics that the Bonferroni correction is "mainly useful when there are a fairly small number of multiple comparisons, and you're looking for one or two that might be significant. However, if you have a large
number of multiple comparisons and you're looking for many that might be significant, the Bonferroni correction may lead to a very high rate of false negatives.” (McDonald, 2014) This analysis fits into the latter category. We have decided not to correct for multiple comparisons.

4.2.2 Principal Component Analysis

We then performed a Principal Component Analysis (PCA) to discern evident clusters or trends visually. We were able to see that much of the variance is captured by the first two principal components (PC 1 and PC 2). A distinct “elbow” occurs at the third Principal component (PC 3). There was no apparent structure visible when we plotted all participants in two-dimensional space defined by the PC1 and PC2 axes.

![Variance Explained by Principal Component](image)

**Figure 7: PCA Variance Explained by Principal Components**

*A plot of variance explained shows distinct “elbow” at the third principal component*
No apparent structure was present when all participants are plotted on the PC1/PC2 axes.

Ten cytokines contribute ~4% each to Principal component 1.
Figure 10: Cytokine Contributions to PC 2

*FGF basic contributes 10% to PC2, while IL17 contributes 9%, and FN.alpha contributes 8%.*
Chapter 5 Model Development

5.1 Baseline binary classification models

After consulting “Supervised Machine Learning, A Review of Classification Methods” (Kotsiantis, 2018) We ran a variety of untuned baseline binary classification models, including all 51 cytokines, and then fine-tuned the best three models. (Kuhn, Building Predictive Models in R, 2008) When partitioning the data into training and test sets, we randomly selected 80% of the data for training, reserving 20% for testing. (Kuhn, Predictive Modeling with R, Data Splitting and Estimating Performance, 2013)

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>0.651</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.669</td>
</tr>
<tr>
<td>SVM Radial</td>
<td>0.661</td>
</tr>
<tr>
<td>SVM Poly</td>
<td>0.675</td>
</tr>
<tr>
<td>LDA</td>
<td>0.653</td>
</tr>
<tr>
<td>KNN</td>
<td>0.597</td>
</tr>
<tr>
<td>GLM Elastic Net</td>
<td>0.663</td>
</tr>
<tr>
<td>Boosted Tree</td>
<td>0.658</td>
</tr>
<tr>
<td>Bagged Tree</td>
<td>0.648</td>
</tr>
<tr>
<td>Neural Net</td>
<td>0.633</td>
</tr>
</tbody>
</table>

Table 2: Baseline Classification Model Accuracy

All baseline models, except for KNN, have accuracy within 2% of 63%. KNN performance was significantly worse
5.2 Classification Model Tuning

5.2.1 Support Vector Machine (Polynomial Kernel)

When using all 51 cytokine values, for the training data set, the best SVM fit has a degree of 3, a scale of 0.01, and a cost coefficient of 1. The baseline model training accuracy was 66.9%.

We then used recursive feature elimination in backward direction to select a feature subset to fine-tune the SVM model. The best cytokine subset was: TGF.beta, CCL2, NGF, IL.15, IL.13, RESISTIN, PAI.1, and G.CSF.

Using that subset of variables, we then iterated upon the model, tuning it, and finding the best fit with a polynomial degree of 3, scale 0.01, and cost 1.
Figure 12: SVM Poly Cost Optimization

*SVM of degree 3 has more accurate performance than degree 1 or 2.*

The final SVM (Polynomial Kernel) model has a test accuracy = 0.607, $\kappa = 0.034$, Sensitivity = 0.171, Specificity = 0.859, Positive Predictive Value (PPV) = 0.412, and Negative Predictive Value (NPV) = 0.642. Of note was the imbalance between Sensitivity and Specificity.
The confusion matrix shows the imbalance between sensitivity and specificity.

The return operator curve (ROC) has an area under the curve (AUC) of 0.542.
Next, we tuned a Random Forest Model, again using backward recursive feature elimination (RFE), to find the best variable subset.

The subset chosen, by the RFE algorithm was: CD40L, TGF.beta, CCL2, IL.1beta, CXCL5, PAI.1, RESISTIN, NGF, IL.15, IL.2, IL.6, CCL5, LEPTIN, PDGF.BB, IL.12p7, and VCAM1.

We employed an exhaustive grid search for the “mtry” parameter (the number of variables available for splitting at each tree node, using the sequence of integers between 1 and 15, thus including all possible numbers. As shown in Figure 14, below, the optimal value for mtry was 6.
The best random forest model has 6 variables available for splitting at each tree node. Its training accuracy was 71.923% and $\kappa = 0.289$. The test accuracy was 0.625, with test $\kappa = 0.078$. Sensitivity = 0.195, Specificity = 0.873, PPV = 0.471, and NPV = 0.653.

![Optimization of Random Forest Parameter 'mtry'](image)

**Figure 16:** RF Model Accuracy vs. Number of Variables Available for Splitting

*When testing mtry between 1 and 15, the peak accuracy occurred at mtry = 6, and mtry = 3.*
An optimized random forest model shows the imbalance between sensitivity and specificity.

Random Forest ROC curve shows the area under the curve of 0.534. Random Forest has a slightly higher AUC than the SVM-Poly model.
5.2.3 Generalized Linear Model (Elastic Net)

We then moved on to tune an elastic net classification model. In this case, we used a stepwise forward selection algorithm to find the best variable subset for elastic net classification. The backward seeking recursive procedure used in the previous two cases failed.

The variable best subset was: CCL5, RESISTIN, TGF.beta, CD40L, PDGF.BB, NGF, CXCL1, IL.2, IFN.gamma, IL.13, IL.8, TNF.beta, CCL3, CCL4, GM.CSF, CCL2, PAI.1, VCAM1, and IL.7.

We then performed an exhaustive grid search for the optimum combination of $\alpha$ and $\lambda$ parameters. The optimized elastic net model has an $\alpha$ coefficient of 0.194, and a $\lambda$ coefficient of 0.0023 with an overall accuracy of 0.75. The area under the ROC was 0.702, while Sensitivity was 0.439, Specificity 0.93, PPV 0.783, and NPV 0.742. Of note, the
balance between PPV and NPV in this model was much better than the previous classification cases.

Figure 20: Optimization of Mixing of $\alpha$ and $\lambda$ Coefficients
The elastic net model has a better balance between sensitivity and specificity, as well as greater accuracy of 75%.

The elastic net model has an AUC of 0.684.
Figure 23: Elastic Net Model ROC (Sensitivity/Specificity)
5.3: Final Binary Classification Model Selection

After tuning the top three binary classification models from the baseline run, each with their best respective variable subsets ascertained through RFE, and their optimized parameters, we compared the results both numerically and graphically.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Area Under ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Poly</td>
<td>0.661</td>
<td>0.542</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.625</td>
<td>0.534</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>0.750</td>
<td>0.703</td>
</tr>
</tbody>
</table>

Table 3: Tuned Model Performance

*When the top three classification models are tuned, the elastic net model far outperforms both the SVM Poly and RF models.*
Figure 24: Tuned Model Performance

The generalized linear model (GLM) elastic net (Enet) was, by far, the best performing binary classification model with statistical significance at the $\alpha=0.05$ level and a McNemar's Test P-Value: 0.001315. We discuss the implications of this finding in the conclusion section.
5.4 Baseline Regression Models

We ran a variety of untuned baseline regression models, including all 51 cytokines. (Kuhn, Building Predictive Models in R, 2008) and then fine-tuned the best three models. We partitioned the data into training and test sets randomly selecting 80% of the data for training and reserving 20% for test. (Kuhn, Predictive Modeling with R, Data Splitting and Estimating Performance, 2013)

All models used 10-fold cross-validation.

The accuracy and $RMSE$ of the first pass models are shown below in Table 3 and Figure 22.

Baseline Untuned Regression Model Metrics

<table>
<thead>
<tr>
<th>Model Name</th>
<th>MAE</th>
<th>RMSE</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boosted Tree</td>
<td>22.154</td>
<td>25.09</td>
<td>0.079</td>
</tr>
<tr>
<td>Linear Reg</td>
<td>22.065</td>
<td>26.057</td>
<td>0.078</td>
</tr>
<tr>
<td>Partial LS</td>
<td>21.751</td>
<td>25.428</td>
<td>0.079</td>
</tr>
<tr>
<td>Lasso</td>
<td>22.289</td>
<td>25.382</td>
<td>0.072</td>
</tr>
<tr>
<td>Random Forest</td>
<td>22.43</td>
<td>25.53</td>
<td>0.054</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>21.91</td>
<td>25.213</td>
<td>0.067</td>
</tr>
<tr>
<td>Bagged Tree</td>
<td>22.01</td>
<td>25.548</td>
<td>0.064</td>
</tr>
<tr>
<td>Neural Net</td>
<td>44.571</td>
<td>51.556</td>
<td>NaN</td>
</tr>
</tbody>
</table>

Table 4: Baseline Regression Model Metrics

The baseline models, except for the artificial neural net, all have mean absolute error ~22%. The artificial neural net model performs poorly, with MAE of 44.57%, and we should discard it.
The baseline untuned performance of all models fitted, except the neural net, was similar. Each of their mean absolute errors hovers around 22%.

The neural net model has 22% more mean absolute error, with an MAE of 44.571, and we should discard the model.

The best three baseline regression models are partial least squares regression, bagged tree regression, and elastic net regression. We tuned each of these three best performing models.
5.5 Regression Model Tuning

5.5.1 Partial Least Squares

When we use all 51 cytokines in a baseline partial least squares model, the best number of components was 3, the RMSE was 25.325, \( R^2 \) was 0.074, and MAE was 21.64.

The best variable subset, as selected by recursive feature elimination, was IFN.gamma, NGF, GM SCF, CXCL1, and SCF.

The best partial least square model optimized number of components at 2 (see figure below), and arrived at a training RMSE of 25.33, a training \( R^2 \) 0.0616, and a training MAE of 22.49. The test RMSE was 25.86, and the test MAE was 23.25.
Figure 26: PLS Regression Model Optimization of Number of Components

The RMSE of the PLS model shows a distinct minimum at two components used.

5.5.2 Bagged Tree Regression

When we use all 51 cytokines in a bagged tree regression model, the training RMSE was 25.62, training R2 was 0.053, and training MAE was 22.14.

The best variable subset, as selected by recursive feature elimination was shown below in Table 5

<table>
<thead>
<tr>
<th>Best Variable Subset (Bagged Tree Regression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGF.beta</td>
</tr>
<tr>
<td>TRAIL</td>
</tr>
<tr>
<td>CCL5</td>
</tr>
<tr>
<td>HGF</td>
</tr>
<tr>
<td>VCAM1</td>
</tr>
<tr>
<td>IL.15</td>
</tr>
<tr>
<td>PDGF.BB</td>
</tr>
<tr>
<td>RESISTIN</td>
</tr>
<tr>
<td>CCL4</td>
</tr>
<tr>
<td>FGF.basic</td>
</tr>
<tr>
<td>IL.4</td>
</tr>
<tr>
<td>VEGF</td>
</tr>
<tr>
<td>IL.1RA</td>
</tr>
<tr>
<td>IL.2</td>
</tr>
<tr>
<td>G.CSF</td>
</tr>
<tr>
<td>TNF.alpha</td>
</tr>
<tr>
<td>CD40L</td>
</tr>
<tr>
<td>ICAM1</td>
</tr>
<tr>
<td>IFN.beta</td>
</tr>
<tr>
<td>PAI.1</td>
</tr>
<tr>
<td>TNF1</td>
</tr>
<tr>
<td>IL.1beta</td>
</tr>
<tr>
<td>CXCL9</td>
</tr>
<tr>
<td>FASL</td>
</tr>
<tr>
<td>LEPTIN</td>
</tr>
<tr>
<td>CCL11</td>
</tr>
<tr>
<td>IL.12p40</td>
</tr>
<tr>
<td>TGF.alpha</td>
</tr>
<tr>
<td>CCL3</td>
</tr>
<tr>
<td>NGF</td>
</tr>
<tr>
<td>IL.1alpha</td>
</tr>
<tr>
<td>CCL2IL&gt;6</td>
</tr>
<tr>
<td>CCL7</td>
</tr>
</tbody>
</table>

Table 5 Bagged Tree Regression Best Variable Subset

The best fit, bagged tree model, has a training RMSE of 25.664, training R^2 of 0.066, and training MAE of 22.139. The test RMSE was 26.66, and the test MAE was 22.13.
This model has slightly lower performance the previously studied partial least squares regression.

5.5.3 Generalized Linear Model Elastic Net Regression

When we use all 51 cytokines as model input, the optimum $\alpha$ coefficient value was 0.55, while the optimized $\lambda$ value was 0.923. This model has a training RMSE of 25.229, a training $R^2$ of 0.069, and training MAE of 21.972.

The best variable subset, as found by recursive feature elimination (RFE), is shown below in Table 6.

<table>
<thead>
<tr>
<th>Best Variable Subset (GLM Elastic Net Regression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCL3</td>
</tr>
<tr>
<td>CCL5</td>
</tr>
<tr>
<td>CXCL1</td>
</tr>
<tr>
<td>GM.CSF</td>
</tr>
<tr>
<td>IFN.gamma</td>
</tr>
<tr>
<td>IL.12p70</td>
</tr>
<tr>
<td>IL.1beta</td>
</tr>
<tr>
<td>IL.5</td>
</tr>
<tr>
<td>IL.7</td>
</tr>
<tr>
<td>NGF</td>
</tr>
<tr>
<td>RESISTIN</td>
</tr>
<tr>
<td>TGF.alpha</td>
</tr>
<tr>
<td>VCAM1</td>
</tr>
</tbody>
</table>

Table 6: GLM Elastic Net Regression Best Variable Subset

The best fit elastic net model has an $\alpha$ value of 0.1 and $\lambda$ of 0.178. The training RMSE was 23.344, training $R^2$ of 0.146, and training MAE of 20.724. The test RMSE was 26.87, and test MAE was 22.69
5.6 Best Regression Model Selection

Tuned Regression Models Mean Absolute Error

<table>
<thead>
<tr>
<th>Model Name</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial Least Squares</td>
<td>21.77</td>
</tr>
<tr>
<td>Bagged Tree</td>
<td>23.36</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>22.57</td>
</tr>
</tbody>
</table>

Table 7: Tuned Regression Models Mean Absolute Error

*Tuned regression models demonstrate that the PLS model has the lowest MAE*
Figure 28: Tuned Regression Model MAE

*Visualization of data shows that the PLS model has the least error*

Using the metric of Mean Absolute Error, the Partial Least Squares model has the best fit. It has an MAE of 22.77.
Chapter 6: Discussion

6.1 Experimental Design

As quoted earlier, in Section 2, Dr. Montoya’s team at Stanford obtained a single serum sample, in 2009, from each of 186 ME/CFS patients and 388 healthy controls. They gathered these samples without any physical, emotional, or neurocognitive stimulation. From each of these samples, the team measured fifty-one serum cytokines. Dr. Montoya matched the ME/CFS patients and the healthy controls by age and sex. The patients and control participants comparable ages, 49.9 and 50.1, respectively. 76.6% of the ME/CFS patients and 77.3% of the control participants were female. A higher proportion of the ME/CFS group, 91.7%, were Caucasian, as compared with 71.2% of the control participants. Ten participants did not answer the question about race. Six of these participants were ME/CFS patients. Four were controls. Dr. Montoya then excluded these ten participants from further analysis. The final sample size was 574.

To join the study, participants had to meet the minimum age of 14. All participants resided in Northern California, so they had access to the Stanford clinic. Each of them signed a Health Insurance Portability and Accountability Act of 1996 (HIPAA) waiver, and an informed consent form. Participants were classified as ME/CFS cases if they met the 1994 CDC CFS case definition (Montoya 2017) Of note, the critical defining symptom of Post Exertional Malaise (PEM) was present in 96.9. Participants were eligible for the control group if they lacked a history of fatigue and did not meet the 1994 CDC case definition. Applicants were excluded from both groups if they had “active or uncontrolled
morbidities that would have interfered with the patient’s ability to participate in the study, particularly conditions or medications causing immunosuppression or immunodeficiency” (Montoya 2017).

Dr. Montoya did not exclude participants due to clinical depression or other mental illnesses. We suspect that these depressed participants were a confounding variable in the analysis process. Lack of mental illness exclusion may explain the moderate and high illness severity scores in the control group.

6.2 Analysis Summary

Upon receipt of the 574-person, 51 cytokines, data set provided by Stanford University; we visualized the data. We promptly found a cohort of participants with illness severity scores > 51/100 within the “healthy” control group. Sick people in the control group confound all analysis of illness severity. We removed these participants via R’s native outlier elimination function.

We next examined a correlation matrix. We confirmed that 10 out of 51 cytokines examined, CCL5, CD40L, CXCL5, IL13, IL2, IL5, PAI1, RESISTIN, and TGF.beta have statistically significant linear correlations with illness severity.

In the original analysis, we successfully fit statistically significant binary classification and regression models for the data set.

Working from Max Kuhn’s guide to Caret (Kuhn 2008), we fit eight candidate binary classification models and selected the top three. We refined each of those three models by performing feature selection via recursive feature elimination, in the case of Random Forest, and a Support Vector Machine with a Polynomial Kernel, and via stepwise algorithm for a Generalized Linear Model Elastic Net.
We compared the tuned Random Forest, Support Vector Machine (Polynomial Kernel), and the Elastic Net models. We concluded that the best fit of the models was the Elastic Net. It has an overall accuracy of 0.75. The ROC was 0.702. Sensitivity was 0.439, and Specificity 0.93. This model was statistically significant at the $\alpha=0.05$ level with a McNemar's Test P-Value of 0.001315.

We also fitted eight regression models with all 51 cytokines and illness severity. The top three of these regression models were Partial Least Squares (PLS), Bagged Tree, and GLM Elastic Net. As above, with classification, we tuned these top three models after performing feature selection.

The best of the regression models was partial least squares. The PLS model has a test RMSE of 25.86, and test MAE of 23.25. The PLS model was an original finding, as it was the first regression study on illness severity.

These findings validate the hypothesis put forth by Drs. Montoya, Klimas, and Broderick that ME/CFS has a distinct serum cytokine biomarker signature. ME/CFS still faces considerable disbelief in the medical community. Models such as this one, which can lead to a confirmatory blood test, can shift the needle on primary care provider buy-in. This type of tool can allow medical care providers to confirm just how sick their severely presenting patients are. Hard numbers such as these will be incredibly useful in the presentation of cases for disability hearings, convincing doctors that patients genuinely are ill, and moving the science on biomarkers forward.

6.3 Future Directions in Research

The data set, as presented, has a scarcity of moderately and severely ill patients. While more data in every class of illness severity would be advantageous, moderately, and
severely ill patients are particularly important. These patient subsets are often excluded from studies as they are generally too sick to attend clinic appointments. Home visits for sample collection would solve this problem.

We were unable to match the Linear Discriminant Analysis (LDA) accuracy of 93% presented by Dr. Broderick for several reasons. (Broderick 2012)

First, Dr. Broderick screened his patient cohort vigorously. He selected a homogeneous subset of ME/CFS patients by sampling adolescents when they immediately upon infection with Infectious Mononucleosis. He followed these patients longitudinally and continued to obtain samples from those who did not recover and did develop ME/CFS.

Second, Dr. Broderick used a 64-cytokines assay, as opposed to Dr. Montoya's 51 cytokines. Each of Dr. Broderick’s Linear Discriminant Analysis (LDA) model variants included the cytokine IL23, with significant variable weighting. IL23 was not present in the Stanford data set we examined. All future research should utilize the most comprehensive cytokine assay available.

Future research directions also include other types of biomarkers. We know that Natural Killer cells, for example, trend low in ME/CFS patients. Clinical findings, such as body temperature, heart rate, and assessment of daily function, can also be added to machine learning diagnostic classification tools. Another useful modeling step would be building classification models to differentiate ME/CFS from other diseases with similar symptomologies, such as Multiple Sclerosis (MS).
Chapter 7: References

Works Cited


Academy of Sciences, 114(34), E7150–E7158. https://doi.org/10.1073/pnas.1710519114


75


Works Consulted


