Differential gene Expression Analysis to distinguish Psoriatic Arthritis in patients with Cutaneous Psoriasis: A Machine Learning Approach

Prof: Dr.Gurjit Randhawa
Jesse J Annear
145429

Conflicts of Interest:

The individual who performed this analysis has a diagnosis of psoriasis and is working with a rheumatologist on a diagnosis of Psoriatic Arthritis. This may make this analysis more prone to bias.

Terminology

PSO - Psoriasis
PSA - Psoriatic arthritis

Background:

Psoriatic arthritis (PSA) is a chronic condition that affects around 30% of individuals with psoriasis, an inflammatory skin disease[1]. It's important to note that 90% of patients with psoriatic arthritis have psoriasis and if they don't have psoriasis there is usually a family member who does, which highlights the genetic nature of psoriasis and psoriatic arthritis. While both psoriasis and PSA worsen over time, diagnosing PSA when psoriasis is present has been a challenge. A study showed that only 23% of patients were diagnosed with PSA at symptom onset, with 45% diagnosed after 2 years[2]. Early diagnosis and treatment are crucial in improving long-term outcomes and preventing joint damage and poor functional outcomes. Additionally, PSA can present similarly to other autoimmune diseases, and according to arthritis.org PSA patients have double the risk of developing cardiovascular disease, 43% more likely to have or develop heart disease and a 23% increased risk of developing conditions that affect blood flow to the brain[3].Dna expression research has led to effective biologic medications like taltz(ixekizumab), which lower the expression of il-17 in patients with psoriasis, PSA, and other diseases. Further research is needed to better define the differences between psoriasis, PSA, and other autoimmune diseases, and to isolate biomarkers that have the potential to revolutionize medical care.

In our analysis, we are exploring the use of gene expression data to develop a machine learning model that can differentiate between psoriatic arthritis and psoriasis(only). We are utilizing differential gene expression analysis to improve the accuracy of our model by reducing the number of dimensions in the data. Developing an accurate diagnostic tool is critical as early courses of biologics have been shown to greatly reduce patient outcomes (management of psoriatic arthritis: early diagnosis, monitoring of disease severity and cutting edge therapies). If a patient can get onto suitable treatment to reduce inflammation early it can help stop aggressive bone, cartilage and tendon damage that otherwise would be irreversible. Identifying specific biomarkers associated with PSA can also help us better understand the condition and develop more targeted treatments. The potential for personalized treatment plans using isolated biomarkers could also be significant. We hope that this analysis will help show the use of gene expression on top of rheumatological diagnosis tools like CASPAR[1]. CASPAR alone may not be ideal as it typically is a lagging indicator as it can only measure inflammatory damage after it has persisted for a good amount of time. That's why gene expression analysis could aid patients in getting a more timely diagnosis and reduce their risks of developing other mental disorders like somatic symptom disorder that commonly goes along with psoriatic patients who take an average of 2.5 years[1] after symptom development to finally get their diagnosis with PSA.

Although some research has been done using differential gene expression analysis in order to identify important biological pathways, there has been little to no research done on

using machine learning models on top of differential gene expression analysis in order to detect psoriatic arthritis in patients with psoriasis. We believe that this kind of analysis may provide a solution to the problem of late diagnosis of psoriatic arthritis. We suggest that the use of an 'all encompassing model' which includes differential gene expression data of multiple cell types, patient medical history, CASPAR scores from a rheumatologist, MRI and ultrasound technology may provide a solution to the late diagnosis problem. Unfortunately, even with the recently popularized use of machine learning and big data, there has been little to no research done on the use of machine learning to aid in the diagnostic process of psoriatic arthritis.

This may be because most of the research done is focused almost exclusively on finding new biological pathways in which biologics can be made to treat and profit immensely from. Currently the biologics space is a multi-billion dollar industry, with it estimated to reach a total industry market value of 719.84 billion by 2023[4]. This can be observed as companies like Abbivie (Stock ticker ABBV) have seen tremendous growth and profits from drugs like humira which can be solidified by constant stock growth due to increasing profit margins and a 'wall of patents' which allowed ABBV to have 165 patents on a single drug, effectively pushing back the opportunity for biosimilars in an effort to 'game' the biologics industry[5]. This isn't to say that biologics aren't important because they are some of the greatest technology advancements that the medical field has observed, but it's important to encourage more research around early diagnosis so patients can start biologics treatments earlier and have improved patient outcomes and reduced health care costs.

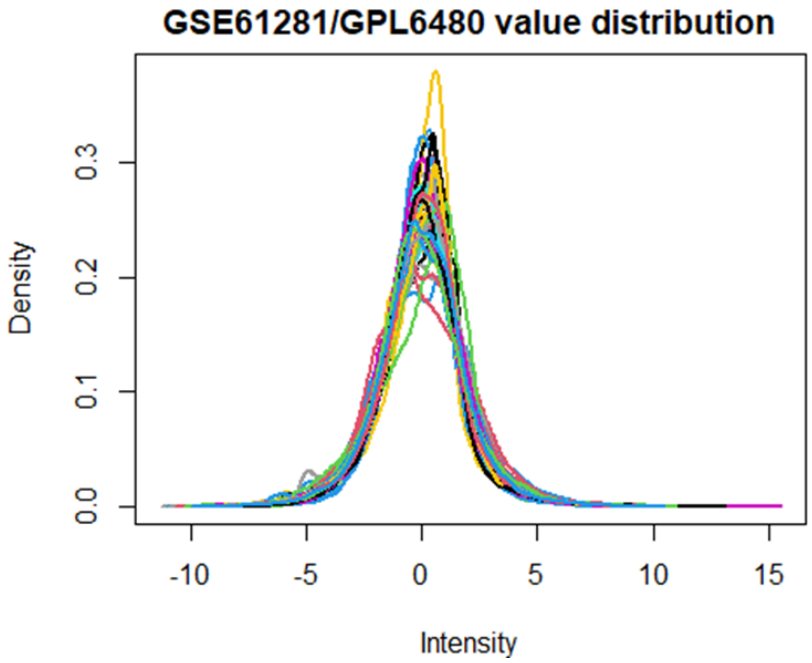
Dataset:

The dataset is public and was posted on September 24, 2014 as a GEO dataset published on the NIH website under GSE61281[6]. It involves human whole blood samples and compares transcriptional profiles between psoriatic arthritis (PSA), cutaneous psoriasis without arthritis (PSO), and unaffected controls. The experiment was conducted using expression profiling by array. The overall design includes three conditions: PSA, PSO, and unaffected controls. There were 20 biological replicates each for PSA and PSO and 12 for controls. Although sampling of inflammatory cells would be more ideal, for example synovial fluid cells as this is a primary area of disease activity, blood cells could provide more information on the high comorbidity rate that is apparent in both psoriasis and psoriatic arthritis patients with cardiovascular diseases and diabetes. On top of an increased rate when compared to controls of these diseases it has also been shown that smoking can increase the chance of developing both of these diseases[7]. So, it is definitely of interest for more studies to be down with cardiovascular specific cells like blood, or even lung cells.

Data exploration:

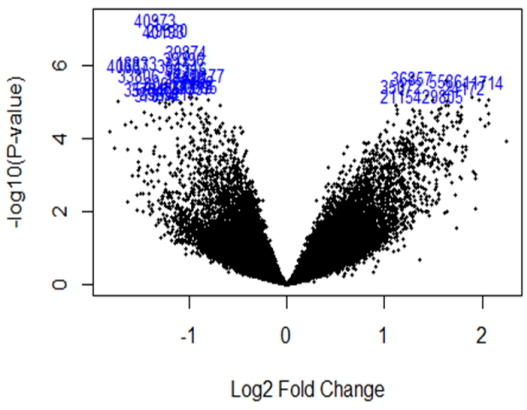
Initially, we conducted an exploratory analysis to identify any significant observations using various visualization techniques. A scatter plot was generated to illustrate the distribution of gene expression values, which indicated a range between -7 and 7. Based on our observations, it appears that the data had been transformed using a log2(x) method. Although it is not always advisable to assume data transformation, log2(x) is a common method used in gene expression data analysis, and the range and distribution of our dataset suggests that it has been log transformed. This is beneficial since it eliminates the need for data scaling and normalization. Furthermore, a density plot was constructed to assess the distribution of

individual patient data, which revealed a relatively consistent distribution amongst patients. This suggests that the data had been previously scaled and normalized before the data was published.



<u>Finding differentially expressed genes:</u>

Differential gene analysis is a powerful tool for gaining insights into the molecular basis of various diseases and disorders, and for identifying genes that are dysregulated in affected individuals. In this study, we utilized the r package limma to isolate genes that were significantly upregulated or downregulated in psoriasis patients compared to controls. We employed the log2(fold change) method to identify significant genes, which is a widely-used technique for detecting differentially expressed genes. We generated a volcano plot to visualize the distribution of our data and identified the top 30 genes that were dysregulated in psoriasis patients.



Surprisingly, our analysis did not identify any genes that were significantly dysregulated in psoriasis patients compared to controls. This finding is unexpected given the increased risk of cardiovascular and metabolic disorders in psoriasis patients. This observation is consistent with a previous study on this dataset that also failed to identify differentially expressed genes between psoriasis patients and controls. The authors of that study suggested that the weak signature of cutaneous disease in whole blood samples might have been obscured

by the heterogeneity of whole blood. However, they also noted that many of the differentially expressed genes between psoriatic arthritis and psoriasis without arthritis had smaller changes in expression in PSO patients compared to controls, suggesting that the more severe phenotype of PSA is an exacerbation of small gene expression changes that occur in PSO.

This finding raises an interesting question about the relationship between psoriasis and psoriatic arthritis. It is possible that psoriatic disease represents a spectrum of disease severity rather than discrete entities, and that the dysregulated genes that we failed to identify in psoriasis patients are still differentially expressed but to a lesser extent. Indeed, some patients with psoriatic arthritis do not show signs of psoriasis, and it is possible that many of these cases go undiagnosed by rheumatologists who do not consider the diagnosis without skin symptoms.

In conclusion, our analysis using limma failed to identify any significantly dysregulated genes in psoriasis patients compared to controls. This unexpected result raises interesting questions about the relationship between psoriasis and psoriatic arthritis and highlights the need for further studies to elucidate the molecular mechanisms underlying these disorders.

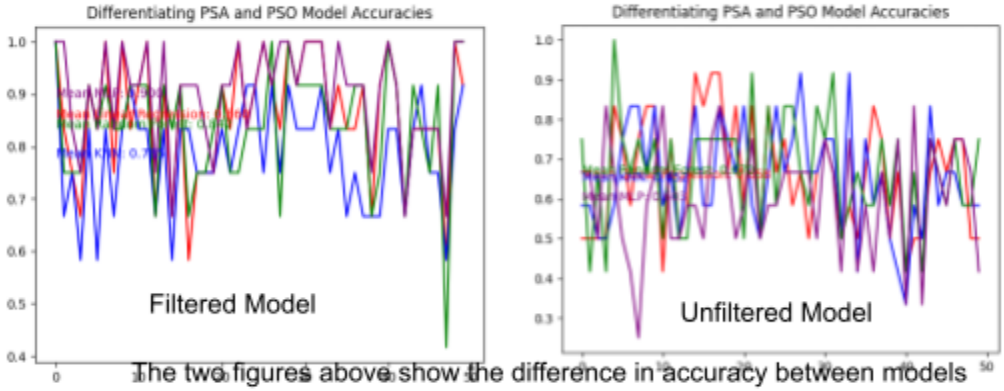
Enrichment Analysis:

Following the identification of significantly expressed genes, the next step in gene expression analysis is to perform enrichment/pathway analysis. Enrichment analysis, also known as gene list analysis, involves comparing and modeling an entire gene list or just the significant genes against a database of various biological pathways and important genetic information. The purpose of this step is to gain a deeper understanding of the biological processes involved in the significant genes or the ordered list of genes sorted by their p-values. In this study, we utilized g: profiler to perform enrichment analysis, and identified several interesting findings related to metabolic disorders that are highly comorbid with psoriasis and psoriatic arthritis. A bubble chart comparing the different categories was generated and revealed significant biologic processes and pathways related to protein kinase camp-activated catalytic subunit alpha, which has been implicated in the development of new target drugs for type 2 diabetes, another common comorbidity associated with psoriatic disease. However, parsing through this vast and complex information requires significant time and biological expertise. Therefore, careful examination and interpretation of the data is necessary to fully appreciate the important insights provided by enrichment analysis.

Modeling phase:

We created a model to differentiate between psoriatic arthritis and psoriasis(only) by identifying significant genes using gene expression data.Since we weren't interested in differentiating between psoriasis and controls we left the controls portion of the data out. In the modeling phase, we tested various machine learning models such as logistic regression, KNN classifier, random forest classifier, and MLP neural network. We found that the MLP neural network had the highest accuracy compared to other models, which was further improved by isolating only the significantly expressed genes. The accuracy of the logistic regression model was measured to be an average of 90%. These results are much better than our model where we did not filter out genes that weren't significant as the MLP only had an accuracy of 67% in that case which is 23% less accurate.Therefore, this shows that differential gene expression analysis is a suitable technique to reduce the dimensionality in our gene expression models

where the data has a high amount of dimensionality and a low amount of data .Our results suggest that the MLP neural network is a suitable method for modeling and classifying psoriatic arthritis in patients with psoriasis. However, to further improve on model efficacy it would be beneficial to combine with information obtained by health care specialists like CASPAR, MRI'S, ultrasounds of tendons and more medical history in order to find a more complete model that can better diagnose psoriatic arthritis in patients with psoriasis. This type of 'all encompassing model' has not been suggested as far as we are concerned and it would likely reduce the time it takes to receive a diagnosis in patients with psoriatic arthritis.



The two figures above show the difference in accuracy between models
Larger images are located under the sources section

Discussion:

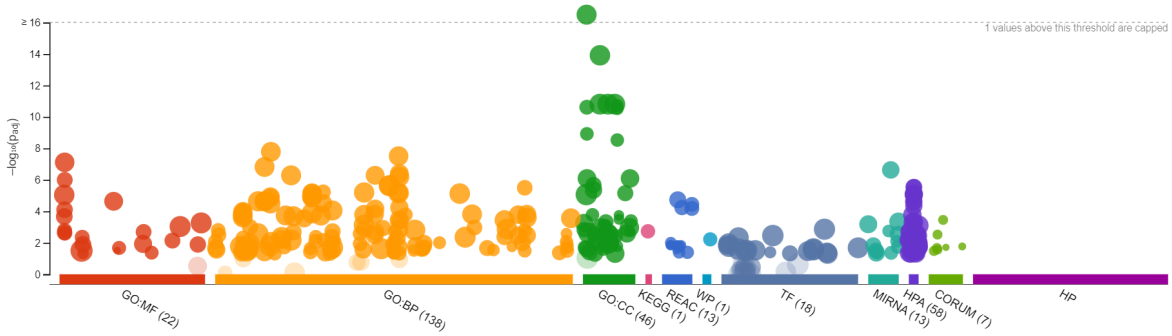
In conclusion, this study aimed to use gene expression data to develop a machine learning model that can differentiate between patients with only psoriasis and those with psoriatic arthritis as well. Early diagnosis of PSA in patients with psoriasis is crucial for improving long-term outcomes and preventing joint damage and poor functional outcomes. The dataset used in this study involved human whole blood samples and compared transcriptional profiles between PSA, cutaneous psoriasis without arthritis (PSO), and unaffected controls. However, the analysis did not identify any genes that were significantly dysregulated in psoriasis patients compared to controls. This finding is unexpected given the increased risk of cardiovascular and metabolic disorders in psoriasis patients. The potential use of isolated biomarkers for personalized treatment plans could be significant, and further research is needed to better define the differences between psoriasis, PSA, and other autoimmune diseases. Although sampling of inflammatory cells would be more ideal, blood cells could provide more information on the high comorbidity rate that is apparent in both psoriasis and PSA patients with cardiovascular diseases and diabetes. Moreover, using an MLP neural network model after identifying significantly expressed genes could potentially aid in the diagnosis of PSA in patients with psoriasis. The model could be used as a tool to predict the likelihood of a patient having PSA based on their gene expression profile. This could potentially help healthcare professionals make more accurate and timely diagnosis, leading to better treatment outcomes for patients.

However, further research is needed to validate the use of such models in clinical settings and likely a larger variety of cell types on top of tests like caspar that a rheumatologist could perform. Further research is needed in order to provide an earlier diagnosis, one where a significant amount of joint damage has occurred. Finding a way to get an earlier diagnosis not only improve PSA patient outcomes but would also lower the burden of a PSA patient on the

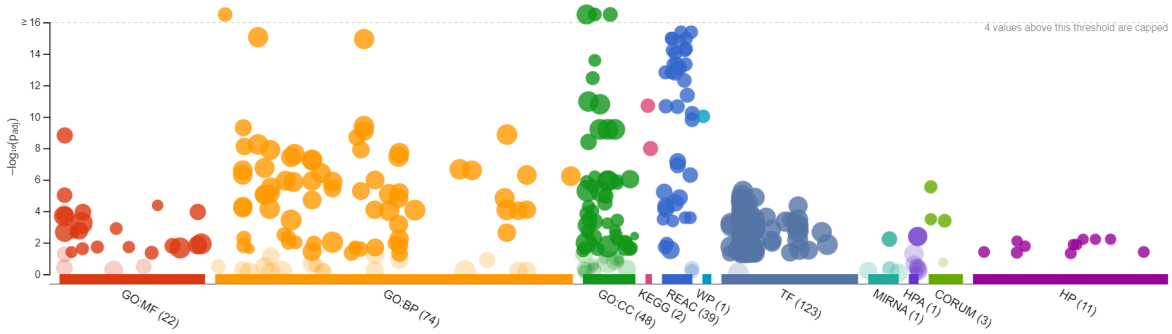
healthcare system by reducing the chance of needing surgery and by having less of a chance to develop comorbidities like metabolic disease, somatic symptom disorder, depression and cardiovascular disease. We suggest that the use of an ‘all encompassing model’ which includes differential gene expression data of multiple cell types, patient medical history, CASPAR scores from a rheumatologist, MRI and ultrasound technology may provide a solution to the late diagnosis problem. Unfortunately, even with the recently popularized use of machine learning and big data, there has been little to no research done on the use of machine learning to aid in the diagnostic process of psoriatic arthritis and we may be the first to suggest such a model to aid in the diagnosing of psoriatic arthritis.

Bonus Enrichment Analysis

> PSO



> PSA



version	e107_eg54_p17_bf42210
date	2022-12-08, 7:27:21 p.m.
organism	hsapiens

g:Profiler

GO:BP		PSA	pso
Term Name	Term ID	p_adj	p_adj
cytoplasmic translation	GO:0002181	1.329×10^{-17}	1.409×10^{-1}
macromolecule biosynthetic process	GO:0009059	4.632×10^{-16}	1.717×10^{-5}
cellular nitrogen compound biosynthetic process	GO:0044271	6.122×10^{-16}	4.730×10^{-6}
cellular biosynthetic process	GO:0044249	2.423×10^{-10}	6.159×10^{-4}
translation	GO:0006412	5.468×10^{-10}	3.816×10^{-2}
organic substance biosynthetic process	GO:1901576	8.430×10^{-10}	1.102×10^{-3}
cellular macromolecule metabolic process	GO:0044260	9.203×10^{-10}	1.007×10^{-4}
peptide biosynthetic process	GO:0043043	2.186×10^{-9}	6.591×10^{-2}
biosynthetic process	GO:0009058	3.629×10^{-9}	2.181×10^{-3}
regulation of macromolecule biosynthetic process	GO:0010556	7.207×10^{-9}	2.162×10^{-5}
peptide metabolic process	GO:0006518	8.495×10^{-9}	8.445×10^{-1}
positive regulation of macromolecule metabolic process	GO:0010604	1.766×10^{-6}	9.685×10^{-9}
regulation of RNA metabolic process	GO:0051252	1.084×10^{-8}	4.145×10^{-7}
amide biosynthetic process	GO:0043604	1.363×10^{-8}	6.778×10^{-2}
positive regulation of nitrogen compound metabolic process	GO:0051173	1.716×10^{-5}	1.685×10^{-8}
regulation of nucleobase-containing compound metabolic ...	GO:0019219	2.146×10^{-8}	3.479×10^{-7}
regulation of nitrogen compound metabolic process	GO:0051171	2.158×10^{-8}	3.605×10^{-6}
protein metabolic process	GO:0019538	3.016×10^{-8}	6.648×10^{-1}
regulation of cellular metabolic process	GO:0031323	3.428×10^{-8}	1.053×10^{-4}
regulation of cellular biosynthetic process	GO:0031326	3.471×10^{-8}	8.250×10^{-6}
positive regulation of metabolic process	GO:0009893	5.099×10^{-6}	9.482×10^{-8}
regulation of biosynthetic process	GO:0009889	1.045×10^{-7}	1.595×10^{-5}
regulation of primary metabolic process	GO:0080090	1.420×10^{-7}	3.965×10^{-6}
DNA-templated transcription	GO:0006351	1.426×10^{-7}	2.412×10^{-4}
nucleic acid-templated transcription	GO:0097659	1.454×10^{-7}	2.437×10^{-4}
positive regulation of RNA metabolic process	GO:0051254	7.220×10^{-6}	1.695×10^{-7}
RNA biosynthetic process	GO:0032774	2.078×10^{-7}	2.969×10^{-4}
positive regulation of nucleobase-containing compound m...	GO:0045935	1.173×10^{-6}	2.354×10^{-7}
regulation of DNA-templated transcription	GO:0006355	2.794×10^{-7}	3.834×10^{-4}
regulation of nucleic acid-templated transcription	GO:1903506	2.847×10^{-7}	3.874×10^{-4}
regulation of RNA biosynthetic process	GO:2001141	3.374×10^{-7}	4.255×10^{-4}
regulation of mRNA metabolic process	GO:1903311	1.000	5.086×10^{-7}
heterocycle biosynthetic process	GO:0018130	6.956×10^{-7}	2.988×10^{-4}
nucleobase-containing compound biosynthetic process	GO:0034654	7.969×10^{-7}	1.451×10^{-4}
aromatic compound biosynthetic process	GO:0019438	8.388×10^{-7}	3.324×10^{-4}
positive regulation of cellular process	GO:0048522	5.514×10^{-6}	1.150×10^{-6}
positive regulation of biological process	GO:0048518	6.511×10^{-5}	1.261×10^{-6}
positive regulation of cellular metabolic process	GO:0031325	1.414×10^{-6}	4.213×10^{-6}
negative regulation of nitrogen compound metabolic process	GO:0051172	3.353×10^{-4}	3.383×10^{-6}

1 to 39 of 39 Page 1 of 1

GO:CC		PSA	pso
Term Name	Term ID	p_adj	p_adj
nucleoplasm	GO:0005654	1.001×10^{-19}	7.499×10^{-19}
cytosolic ribosome	GO:0022626	2.756×10^{-19}	1.000
ribosomal subunit	GO:0044391	3.759×10^{-19}	1.000
nuclear lumen	GO:0031981	9.747×10^{-12}	9.113×10^{-15}
cytosolic large ribosomal subunit	GO:0022625	2.914×10^{-14}	1.000
large ribosomal subunit	GO:0015934	3.788×10^{-13}	1.000
spliceosomal complex	GO:0005681	1.538×10^{-4}	2.118×10^{-12}
organelle lumen	GO:0043233	4.181×10^{-10}	1.325×10^{-11}
membrane-enclosed lumen	GO:0031974	4.181×10^{-10}	1.325×10^{-11}
intracellular organelle lumen	GO:0070013	4.181×10^{-10}	1.325×10^{-11}
cytosol	GO:0005829	1.592×10^{-11}	1.804×10^{-3}
U2-type precatalytic spliceosome	GO:0071005	1.179×10^{-6}	2.159×10^{-11}
precatalytic spliceosome	GO:0071011	1.838×10^{-6}	3.447×10^{-11}
catalytic step 2 spliceosome	GO:0071013	4.039×10^{-4}	1.535×10^{-10}
U2-type spliceosomal complex	GO:0005684	1.437×10^{-6}	1.377×10^{-9}
ribosome	GO:0005840	4.644×10^{-9}	1.000
catalytic complex	GO:1902494	1.042×10^{-6}	3.339×10^{-7}
cytosolic small ribosomal subunit	GO:0022627	6.711×10^{-7}	1.000

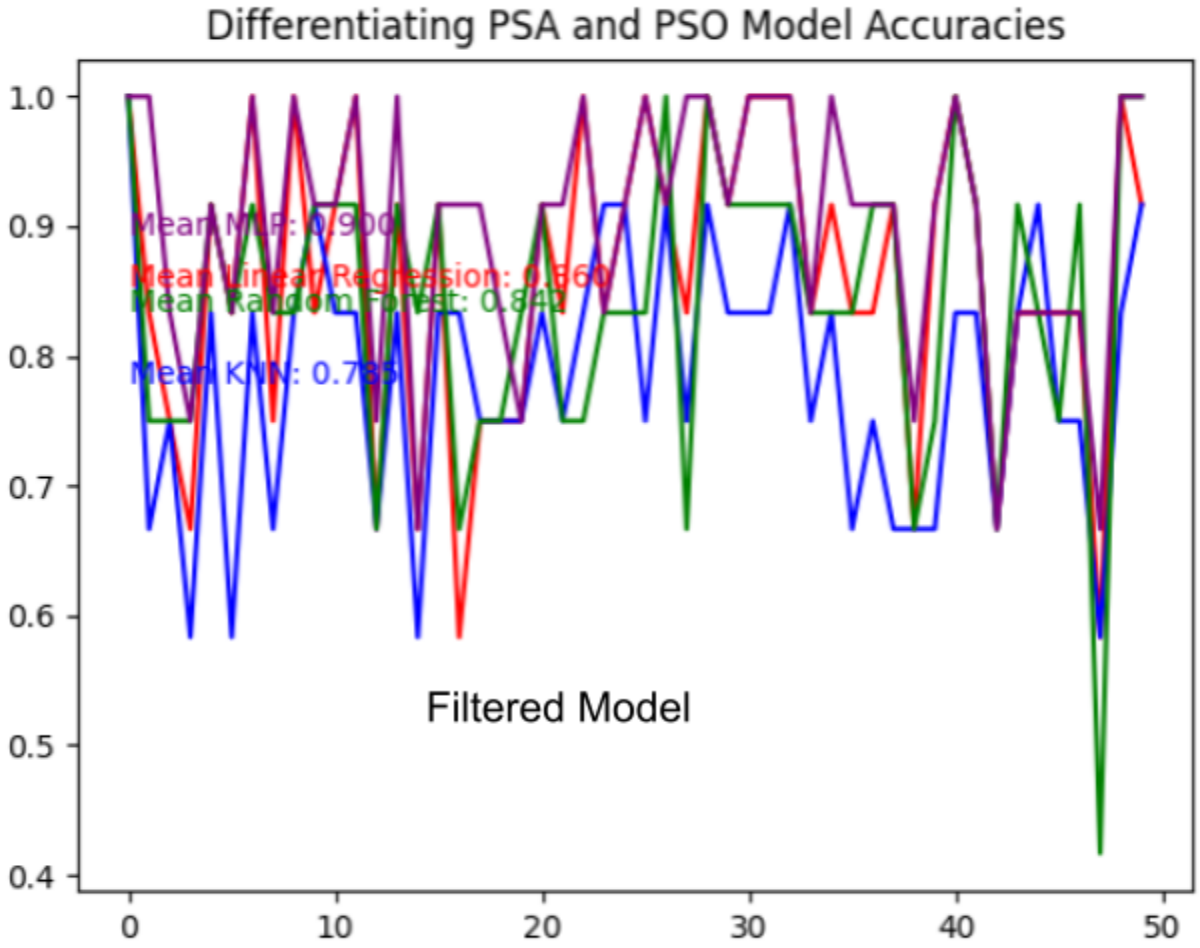
1 to 18 of 18 Page 1 of 1

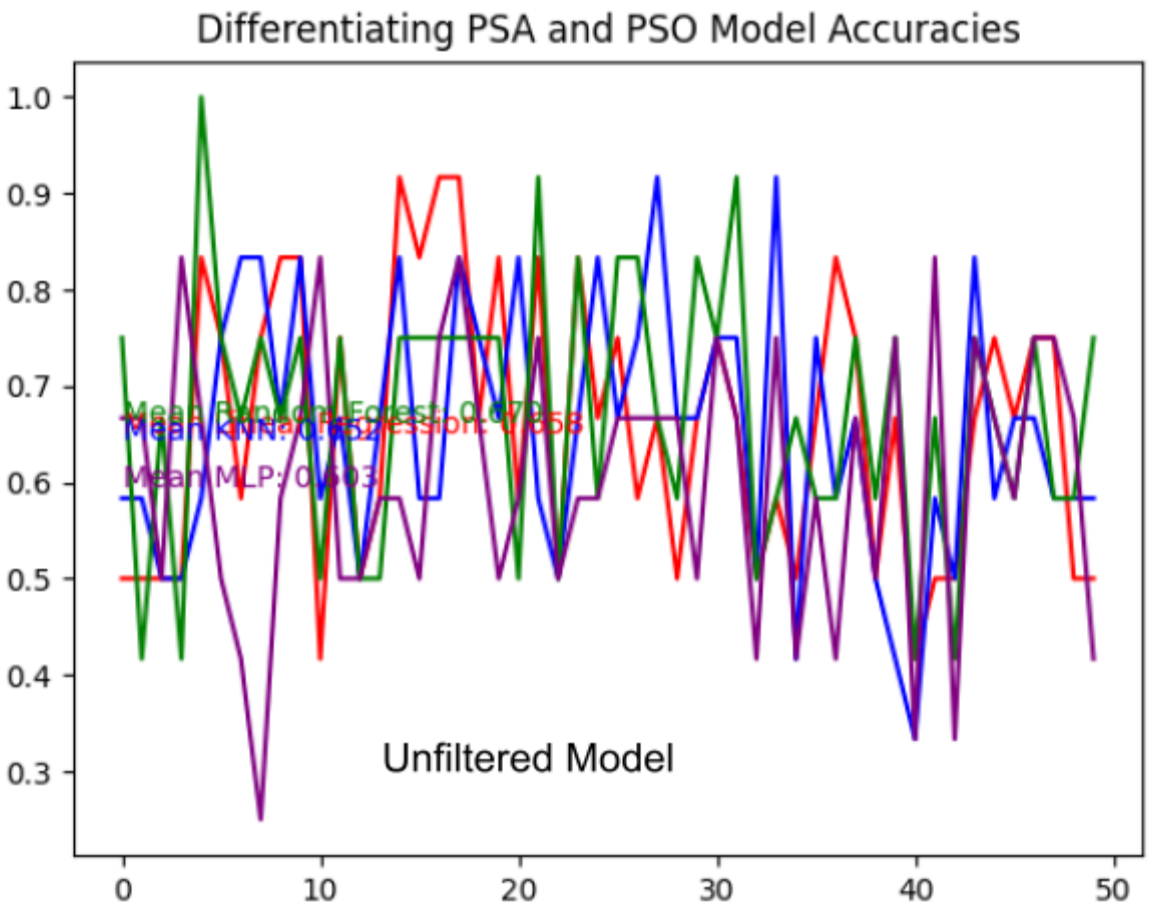
GO:MF		PSA	PSO
Term Name	Term ID	p_adj	p_adj
structural constituent of ribosome	GO:0003735	1.538×10^{-9}	1.000
RNA binding	GO:0003723	2.165×10^{-3}	7.685×10^{-8}
mRNA binding	GO:0003729	5.425×10^{-2}	1.006×10^{-6}
nucleic acid binding	GO:0003676	2.009×10^{-4}	9.039×10^{-6}
transcription coactivator activity	GO:0003713	9.616×10^{-6}	2.276×10^{-3}
enzyme binding	GO:0019899	5.531×10^{-1}	2.300×10^{-5}
mRNA 5'-UTR binding	GO:0048027	4.265×10^{-5}	1.000
transcription coregulator activity	GO:0003712	1.782×10^{-4}	7.904×10^{-5}
DNA-binding transcription factor binding	GO:0140297	1.117×10^{-4}	1.282×10^{-2}
transcription factor binding	GO:0008134	1.152×10^{-4}	1.263×10^{-2}
chromatin binding	GO:0003682	3.863×10^{-1}	2.153×10^{-4}
heterocyclic compound binding	GO:1901363	1.196×10^{-2}	5.451×10^{-4}
protein binding	GO:0005515	5.696×10^{-4}	3.166×10^{-2}
organic cyclic compound binding	GO:0097159	2.183×10^{-2}	9.270×10^{-4}
nuclear receptor coactivator activity	GO:0030374	1.277×10^{-3}	1.000
cAMP-dependent protein kinase activity	GO:0004691	1.587×10^{-3}	1.000
transcription corepressor activity	GO:0003714	1.000	1.633×10^{-3}
structural molecule activity	GO:0005198	1.735×10^{-3}	1.000
cadherin binding	GO:0045296	3.313×10^{-1}	2.034×10^{-3}
mRNA 3'-UTR binding	GO:0003730	1.000	2.605×10^{-3}
cyclic nucleotide-dependent protein kinase activity	GO:0004690	2.720×10^{-3}	1.000
protein C-terminus binding	GO:0008022	1.000	4.451×10^{-3}
RNA polymerase II-specific DNA-binding transcription facto...	GO:0061629	1.615×10^{-2}	7.239×10^{-3}
protein-containing complex binding	GO:0044877	1.000	1.184×10^{-2}
transcription regulator activity	GO:0140110	1.434×10^{-2}	3.009×10^{-1}
oxidoreduction-driven active transmembrane transporter ac...	GO:0015453	1.909×10^{-2}	1.000
peptide N-acetyltransferase activity	GO:0034212	1.916×10^{-2}	1.000
chromatin DNA binding	GO:0031490	1.000	2.123×10^{-2}
N-acetyltransferase activity	GO:0008080	2.419×10^{-2}	1.000
U5 snRNA binding	GO:0030623	1.000	3.056×10^{-2}
histone acetyltransferase activity	GO:0004402	3.993×10^{-2}	1.000
protein N-terminus binding	GO:0047485	4.411×10^{-2}	4.253×10^{-2}
poly(A) binding	GO:0008143	1.000	4.778×10^{-2}

1 to 33 of 33 Page 1 of 1

Sources

1. Psoriatic Arthritis | NEJM
2. Diagnostic Delay in Psoriatic Arthritis: A Population-based Study | The Journal of Rheumatology (jrheum.org)
3. Psoriatic Arthritis and Your Heart
4. Biologics Market Size to Worth Around US$ 719.84 Bn by (globenewswire.com)
5. How a Drug Company Made $114 Billion by Gaming the U.S. Patent System - The New York Times (nytimes.com)
6. GEO Accession viewer (nih.gov)
7. The relationship between smoking, psoriasis and psoriatic arthritis: Expert Review of Clinical Immunology: Vol 15, No 1 (tandfonline.com)


Differentiating PSA and PSO Model Accuracies
Mean MLP: 0.900
Mean Linear Regression: 0.860
Mean Random Forest: 0.842
Mean KNN: 0.735
Filtered Model
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
10
20
30
40
50


Differentiating PSA and PSO Model Accuracies
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
10
20
30
40
50
Unfiltered Model