Dataset: GSE60424
This dataset explores transcriptomic signatures of six immune cell subsets and whole blood samples from patients with various immune-associated diseases, including sepsis, type 1 diabetes, amyotrophic lateral sclerosis (ALS), and multiple sclerosis (MS). The study also includes healthy controls.
Study design: RNA was extracted from:

1. Immune Cell Subsets: Neutrophils, monocytes, B cells, CD4 T cells, CD8 T cells, and natural killer (NK) cells.
2. Whole Blood Samples: Taken from patients and healthy subjects.

Platform: RNA-seq libraries were prepared using Illumina TruSeq.
Sequencing was performed on an Illumina HiScan platform with a target read depth of ~20 million reads per sample.
By filtering the sepsis patients and healthy control ones the dataset is prepared for downstream analysis.(GSE60424code.R)

(sepsis_dataGSE60424.csv) -> sepsis 20 and healthy 28. (47 rows and 57 columns)

Present genes : 55
Missing genes: 0

The data is already normalized and ready for downstream analyses, such as differential expression or machine learning models. Count data were normalized using the **TMM (Trimmed Mean of M-values)** method with the **edgeR package**.

Random Forest

Random forest Model:

The random forest test is done on sepsis_dataGSE60424.csv) dataset and the target label was considered as factors(Sepsis and healthy). (RF-factorGSE60424code.R)

here is the results:

average metrics results:

After 100 repeated random forest by randomly splitting the data (repeated_splits_metrics60424.csv), the average metrics was :(average_metrics604242.csv)

"MCC"	F1	AUC	TPR	TNR	PPV	NPV
0.547929903837531	0.65990315526398	0.83725	0.5475	0.912	0.869214285714286	0.725833333333333

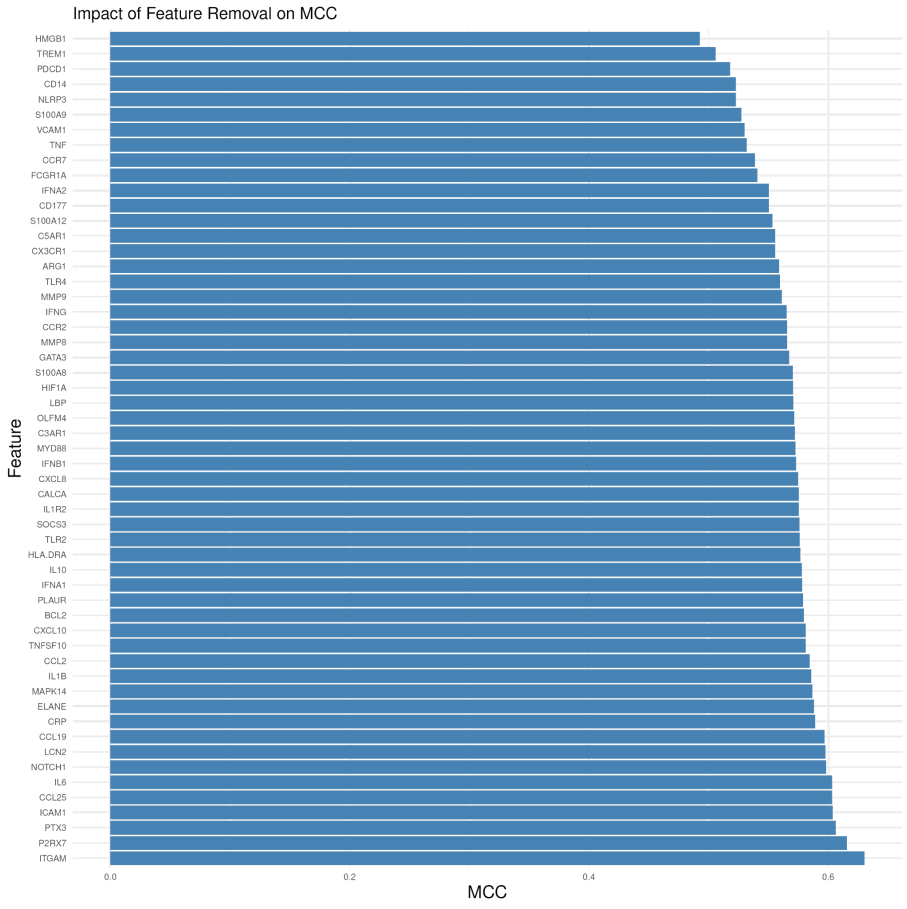
Feature removal random forest:

Shows by removing which column that represents one gene of interest the performance for prediction will drop significantly.(feature_removal_results604242.csv)

feature removing plot:

X-Axis: The values along the x-axis represent the performance difference (using MCC) when a specific feature is removed from the model.

Y-Axis: The features (genes) are listed. Features at the top show a greater drop in performance upon removal, while features at the bottom show a smaller drop in performance.



Top Features (HMGB1, TREM1, PDCD1, CD14, NLRP3):

These are likely key biomarkers and should be prioritized in biological interpretation and downstream analysis. They are central to the model's predictions and could be explored further for biological relevance.

Bottom Features (ITGAM, P2RX7, PTX3, ICAM1, CCL25):

These features might not be crucial for prediction in this context. Consider removing them to simplify the model or explore their redundancy with other features.

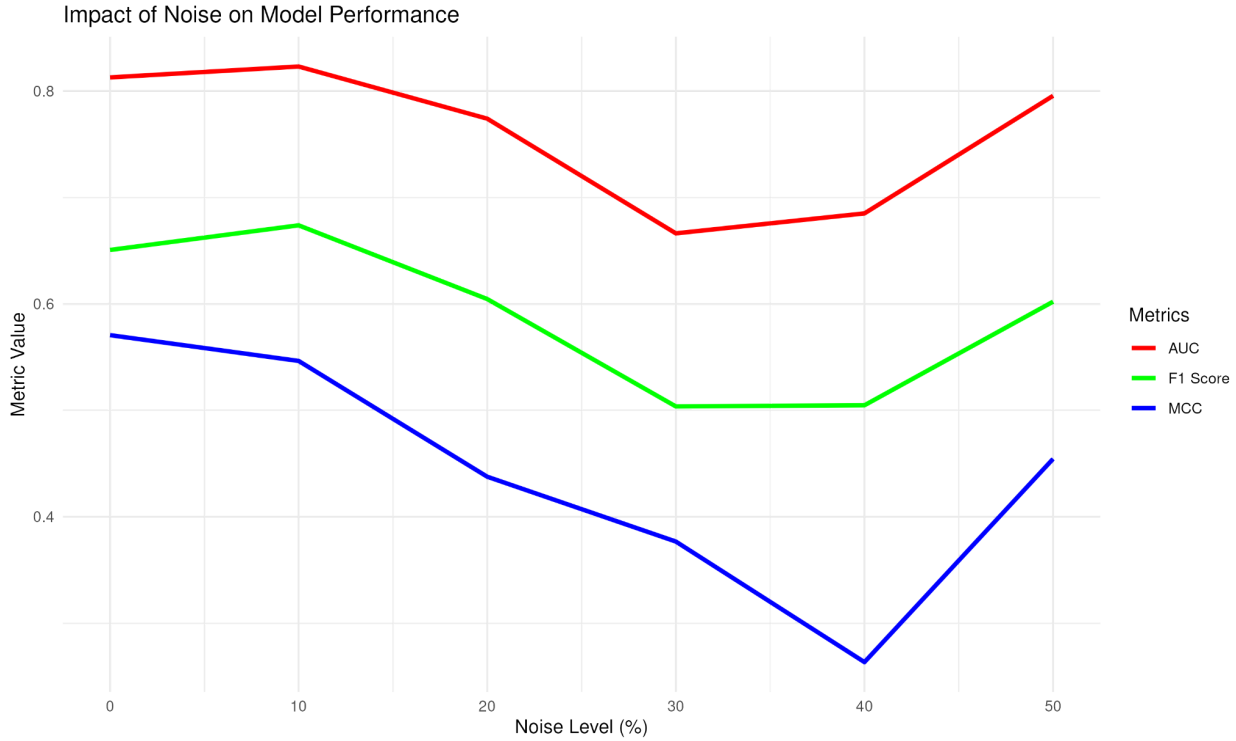
Sanity check

Noise is artificially introduced into the dataset, and the model's performance metrics are observed as the noise level increases.(sanity_check_results60424.csv)

Noise Level 0: Represents the original data with no added noise (baseline performance).

Noise Levels 10–50: Increasing levels of noise are added to the input data, simulating scenarios with reduced data quality or signal interference.

Impact of Noise:



Moderate Noise (10–20): The model shows minor declines in MCC, F1, and AUC.

High Noise (30–40): Significant degradation in performance, particularly in MCC, AUC, and F1. This suggests the model struggles with highly noisy conditions.

There is a problem with the many features or genes columns in the dataset that have 0 values. I do not know if I have to keep them in the dataset because they have meaningful value of expression or they are just redundants and useless.

To investigate it i tried to perform feature importance analysis and here was the results:

```
ordered_importance
           healthy      Sepsis MeanDecreaseAccuracy MeanDecreaseGini Feature
NOTCH1   2.9618201  5.33890973            5.2479512       1.19684872  NOTCH1
CD177    6.6139025  5.85406751            7.2775431       1.17998334   CD177
S100A12  4.9842005  3.45793047            5.3290241       1.02122659 S100A12
ITGAM    4.1818815  3.28744131            4.9241645       0.83244837   ITGAM
MMP9     4.0776075  4.10852692            4.7034047       0.82152981    MMP9
C3AR1    3.6778032  4.03831008            4.1914125       0.79901247   C3AR1
MAPK14   2.4110556  2.10921232            2.7700711       0.65051078  MAPK14
S100A8   3.3305658 -0.24397838            2.5400176       0.59331541  S100A8
MMP8     2.6025316  2.39068335            3.3496277       0.59280068    MMP8
ARG1     2.4456568  3.11455171            3.4837851       0.59020901    ARG1
MYD88    1.3166771  0.36915068            1.1950989       0.53131234   MYD88
IL10     2.5618187  2.92519120            3.4884442       0.49446740    IL10
ICAM1   -0.3976921  2.02510663            1.0570289       0.46629588   ICAM1
IL1R2    3.6477508  0.81013285            3.0134145       0.44932291   IL1R2
HLA.DRA  1.5474109  0.27509239            1.0414182       0.41285947 HLA.DRA
S100A9   0.6865138  0.21027657            0.9812328       0.39676249  S100A9
TNF     -1.2712649  0.62752144           -1.0473169       0.39036479     TNF
CXCL8    2.9998508 -0.74510423            1.6935746       0.36672483   CXCL8
HMGB1    0.2277114 -1.60993839           -0.5363027       0.36415961   HMGB1
LCN2     1.5523774 -1.17146142            0.5400095       0.33631087    LCN2
TLR2    -1.3794092  0.30066511           -0.9419157       0.33380549    TLR2
PDCD1   -0.4634596  0.15553859           -0.3360622       0.31659344   PDCD1
P2RX7    0.9623999  1.50681105            1.7584232       0.29476354   P2RX7
BCL2    -0.3918921 -0.95555292           -0.7862170       0.29130846    BCL2
CCR2    -0.8099792 -0.44781511           -0.7575186       0.29003898    CCR2
CCR7     0.7889048 -0.60729821            0.5885973       0.27394261    CCR7
IL1B     1.7961636 -1.63230825            0.7275856       0.26888233    IL1B
TNFSF10  2.0111612 -0.06462981            1.4976415       0.26478773 TNFSF10
OLFM4    1.2795178  1.16552718            1.6373230       0.26287301   OLFM4
CX3CR1   1.0982040 -2.62046317           -1.0684915       0.25980083  CX3CR1
TREM1    1.6656962 -1.00887654            1.0810641       0.25266261   TREM1
CXCL10   2.7142426  1.23857596            2.4473040       0.24669340  CXCL10
FCGR1A   1.6871212  1.72254929            2.4119704       0.24284238  FCGR1A
NLRP3   -0.4421708 -0.56234044           -0.5209899       0.23628740   NLRP3
SOCS3   -2.2065490 -1.42475110           -2.5909976       0.23348822   SOCS3
HIF1A   -1.5704574 -1.28465871           -1.8191937       0.21274531   HIF1A
```

```
C5AR1    0.3916386 -3.94279613          -1.8298212       0.20432637   C5AR1
PLAUR    1.6987427 -0.18990720           0.9186873       0.20042514   PLAUR
TLR4     1.3320340 -1.75877991          -0.3451813       0.19963233    TLR4
CD14    -0.9446667 -1.07003168          -1.3686245       0.18672605    CD14
ELANE    2.2597540 -2.89993105          -0.6252729       0.18109764   ELANE
IFNG    -1.4887608 -0.91080829          -1.3748363       0.15128576    IFNG
GATA3   -0.2532456 -0.46302867          -0.8500052       0.15034958   GATA3
IL6     -2.0015031 -1.86868033          -2.6555909       0.12825388     IL6
PTX3    -0.0379873 -2.25475237          -1.4646000       0.12819407    PTX3
CCL2    -1.5037410 -0.79018268          -1.8905718       0.06976637    CCL2
CCL25   -1.3783671 -1.00100150          -1.7310546       0.03357742   CCL25
VCAM1    0.0000000  0.00000000           0.0000000       0.01581974   VCAM1
CALCA    0.0000000  0.00000000           0.0000000       0.00000000   CALCA
LBP      0.0000000  0.00000000           0.0000000       0.00000000     LBP
CRP      0.0000000  0.00000000           0.0000000       0.00000000     CRP
IFNB1    0.0000000  0.00000000           0.0000000       0.00000000   IFNB1
CCL19    0.0000000  0.00000000           0.0000000       0.00000000   CCL19
IFNA2    0.0000000  0.00000000           0.0000000       0.00000000   IFNA2
IFNA1    0.0000000  0.00000000           0.0000000       0.00000000   IFNA1
```

We can see the features with zero values are at the bottom of the list.
Then i tried to t-test for one of the features:

```
t.test(data$LBP ~ data$Label)

       Welch Two Sample t-test

data:  data$LBP by data$Label
t = NaN, df = NaN, p-value = NA
alternative hypothesis: true difference in means between group healthy and group
Sepsis is not equal to 0
95 percent confidence interval:
 NaN NaN
sample estimates:
mean in group healthy  mean in group Sepsis

                    0                     0
```

Based on these results I decided to apply random forest again to data while removing features with 0 values.(RF-remove0valuesGSE60424code.R)

Random forest after removing zero value features:

average metrics results:

After 100 repeated random forest by randomly splitting the data (repeated_splits_metrics60424.csv), the average metrics was :(average_metrics604242.csv)

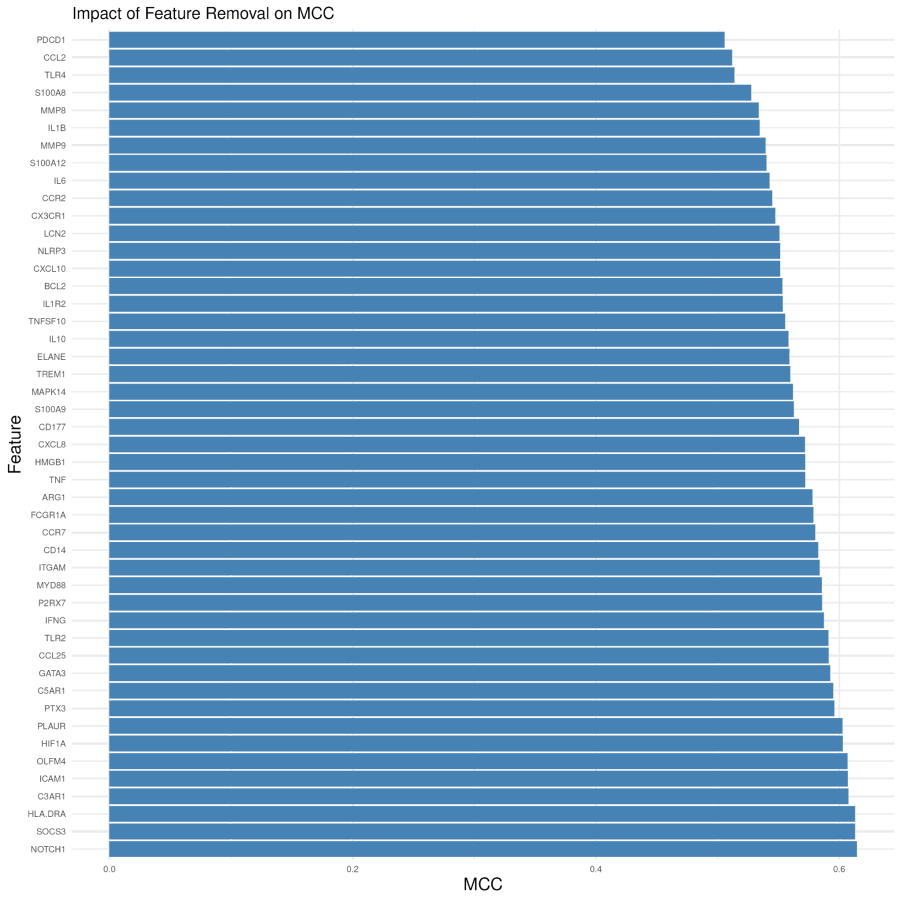
"MCC"	F1	AUC	TPR	TNR	PPV	NPV
0.600066874756967	0.693354094884707	0.834	0.57	0.946	0.918214285714286	0.742369047619048

When we compare the results of 2 average metrics after removing features with 0 values, we can see there are some improvements in all metrics.
Removing the features with all zero values likely reduced noise or irrelevant information in the dataset, allowing the model to focus on meaningful predictors. These zero-value features probably added little to no predictive power and may have been detrimental due to unnecessary complexity.

Feature removal random forest:

Shows by removing which column that represents one gene of interest the performance for prediction will drop significantly.(feature_removal_results604242.csv)

feature removing plot:



Top Features (CCL2, TLR4, PDCD1, S100A8, MMP8):

These are likely key biomarkers and should be prioritized in biological interpretation and downstream analysis. They are central to the model's predictions and could be explored further for biological relevance.

Bottom Features (HLA-DRA, SOCS3, NOTCH1, ICAM1, OLFM4):

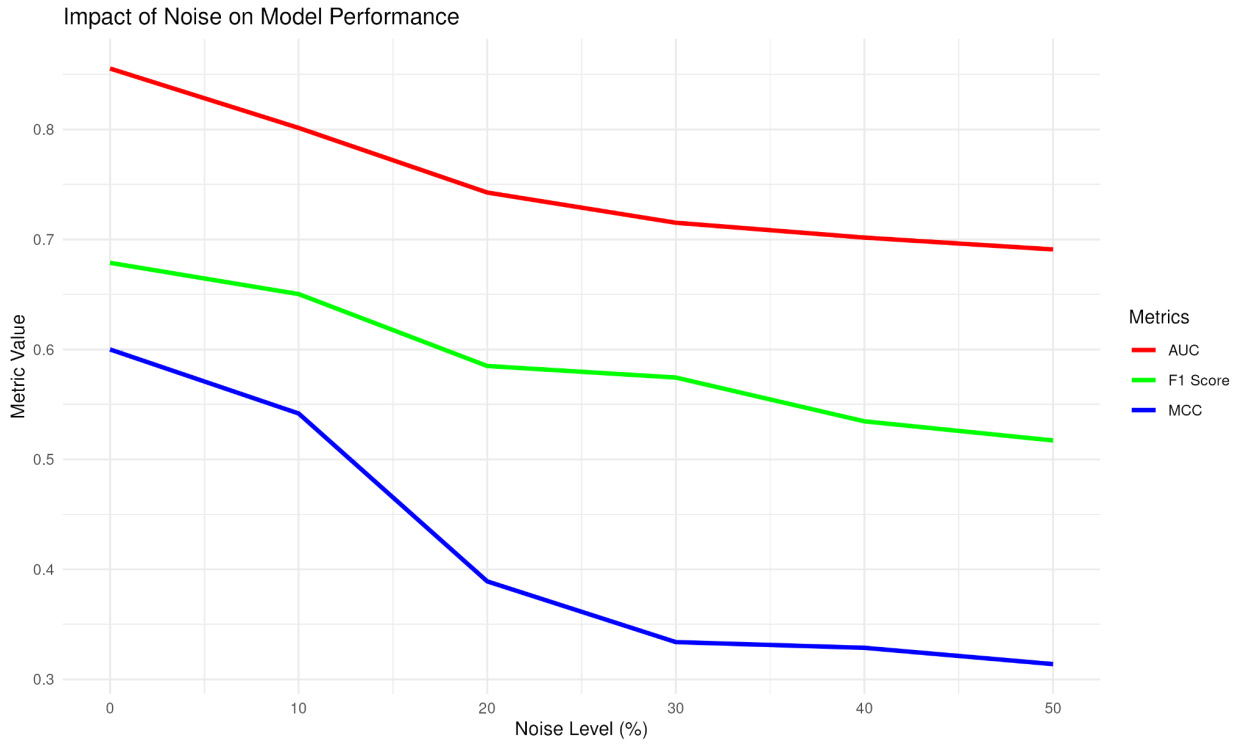
These features might not be crucial for prediction in this context. Consider removing them to simplify the model or explore their redundancy with other features.

Sanity check

Noise is artificially introduced into the dataset, and the model's performance metrics are observed as the noise level increases.(sanity_check_results60424.csv)

Noise Level 0: Represents the original data with no added noise (baseline performance).

Noise Levels 10–50: Increasing levels of noise are added to the input data, simulating scenarios with reduced data quality or signal interference.



Noise Impact:

The addition of noise reduces the model's performance on all metrics, which is expected.

At lower noise levels (0-10%), the model performs well across all metrics, showing only slight degradation.

At higher noise levels (40-50%), the performance deteriorates considerably, especially for MCC and F1 Score, which are more sensitive to misclassifications.