Dataset Description: GSE154918

This dataset, publicly available on GEO (Gene Expression Omnibus), is titled **"RNA-seq analysis of blood from sepsis patients and healthy controls"**. It features RNA sequencing data collected from **peripheral blood samples** of patients, enabling the study of gene expression changes between various patient groups. The study aimed to identify distinct transcriptional signatures associated with the following groups:

- Healthy control (Hlty)
- Uncomplicated infection (Inf1_P)
- Sepsis (Seps_P)
- Septic shock (Shock_P)
- Follow-up of sepsis (Seps_FU)
- Follow-up of septic shock (Shock_FU)

The RNA-seq data were generated using the **Illumina HiSeq 4000 platform**.

Data Processing: In this analysis, the primary focus was on the **sepsis** and **healthy control** groups, along with the **septic shock** group for comparative studies. The following subsets of the data were created:

1. **Sepsis vs. Healthy Control**: A subset was created to focus on the transcriptional differences between sepsis patients and healthy controls. (sepsis_only_labeled) -> 40 healthy and 20 sepsis samples.

2. **Sepsis vs. Septic Shock and Healthy control**: Another subset was created to explore the differences in gene expression between sepsis and septic shock patients.(sepsis_shock_labeled) -> 40 healthy , 20 sepsis and 19 septic shock samples.

3. **Septic Shock vs. Healthy Control**: This subset compares septic shock patients with healthy controls to identify unique gene expression patterns associated with the most severe cases.(shock_only_labeled) -> 40 healthy and 19 septic shock.

4. **Septic and Healthy control** but here consider all the septic shock and sepsis patients as sepsis.(all_sepsis_labeled) -> 40 healthy and 30 sepsis patients samples.

These subsets were prepared by filtering the status.ch1 column of the dataset based on the corresponding group identifiers (Hlty, Seps_P, and Shock_P). Data was normalized by Deseq2 in the database.

Presented-genes: 54

Missing-genes :0

Dataset had already been normalized. This is the result for **only sepsis and healthy controls**.(sepsis_only_labeledGSE154918.csv)

Random forest model:

The random forest is done by considering the target label as a factor.

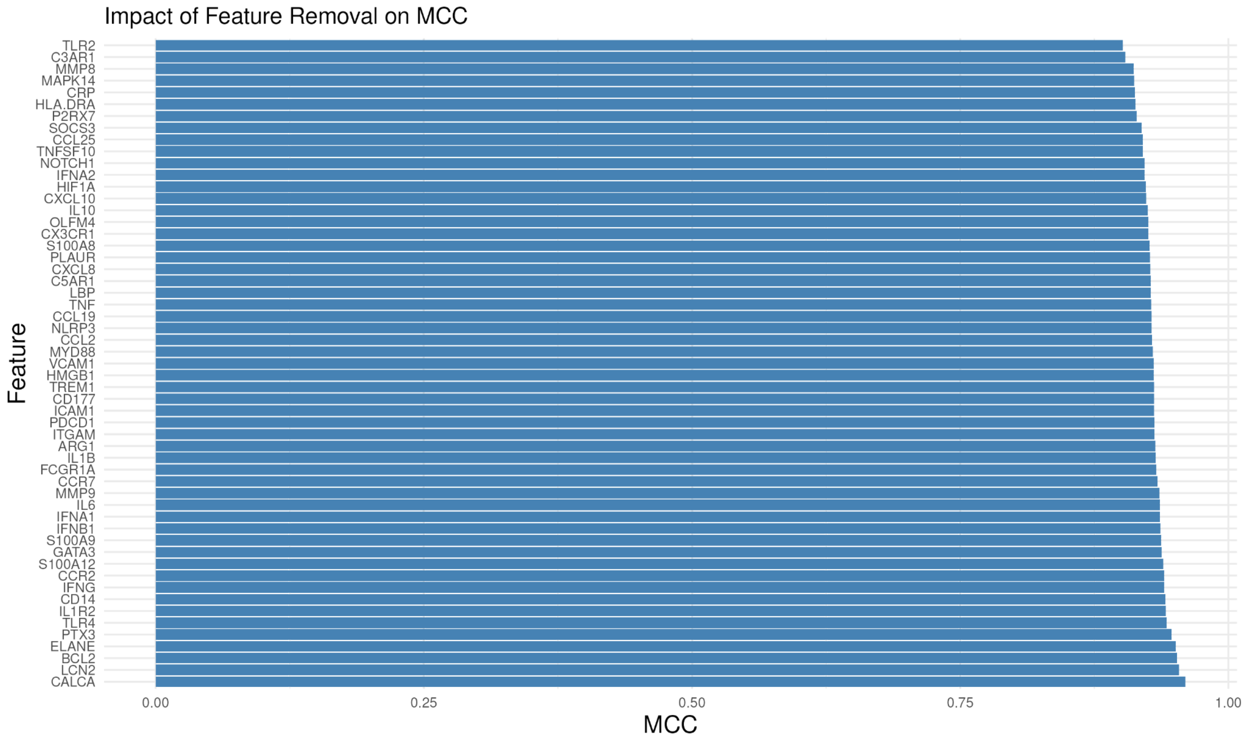
(RFlabel-seponlyGSE154918code.R).

Average metrics values : (average_metrics.csv)

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.9434903	0.9539683	0.9982813	0.9275000	0.9975000	0.9960000	0.9688889

Feature Removal Impact (Feature Importance Plot):

The feature removal plot shows **Confidence:** Provides reassurance that our model's results are grounded in sound logic, not artifacts of the process or data.(feature_removal_results.csv)

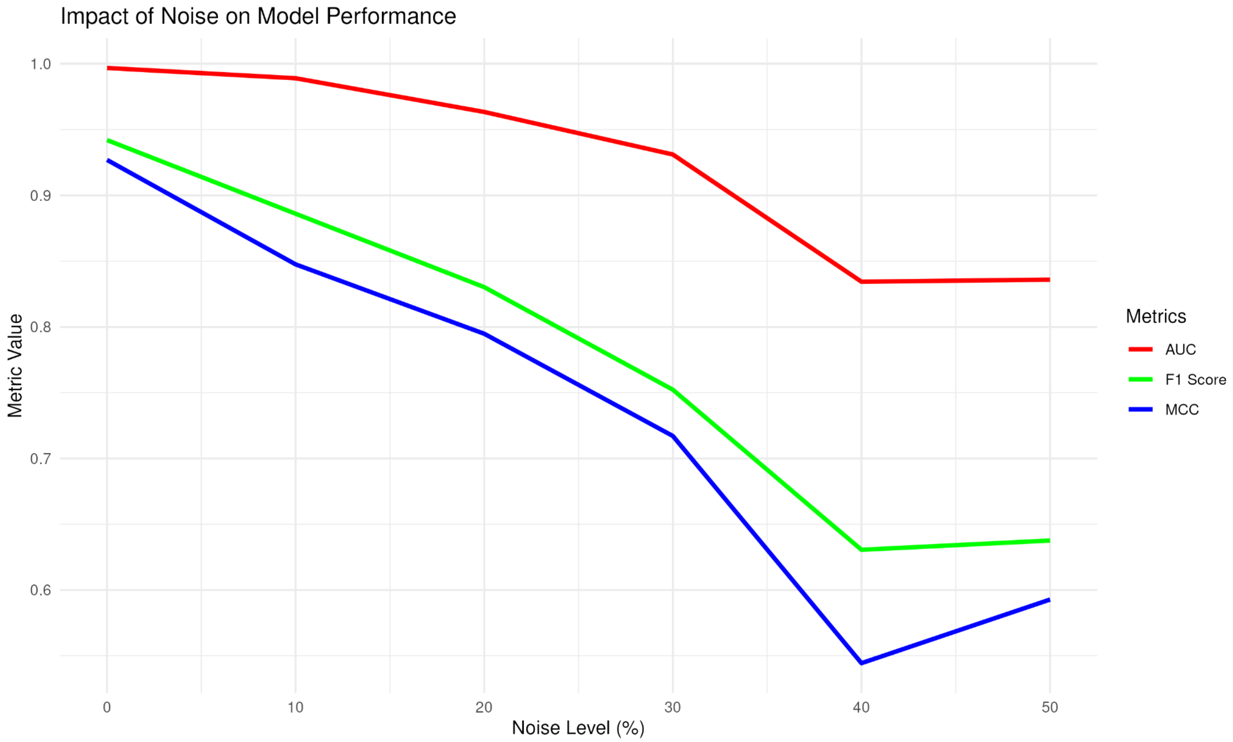


Based on plot analysis, the genes at the top (e.g., TLR2, C3AR1, MMP8, MAPK14 and CRP) cause larger differences in performance (e.g., MCC) when removed, while the genes at the bottom (e.g., CALCA, LCN2, ELANE, BCL2, PTX3) cause smaller differences in performance when removed.

The **genes at the top of the plot** (e.g., **TLR2, C3AR1, MMP8**) are **more important** for the model's predictions because their removal results in a larger impact on performance.

The **genes at the bottom of the plot** (e.g., **CALCA, LCN2, ELANE, BCL2**) are **less important**, as their removal causes minimal impact on performance.

Sanity check(Impact of Noise on Model Performance):(sanity_check_results.csv)
Performance degrades significantly beyond a certain noise threshold (~40 on the x-axis). This indicates the model's reliance on signal integrity.



Gradual Decline: The noise plot shows a gradual decline in performance metrics as noise increases. This is expected behavior and indicates the model is robust to small perturbations.
Sharp Drop Beyond Threshold: The sharp decline after a certain noise level (~40 on the x-axis) suggests the model becomes unreliable when the signal-to-noise ratio is too low.