Dataset: **GSE13015**

This dataset, **GSE13015**, focuses on identifying a blood biomarker signature for the diagnosis of septicemic melioidosis, which is caused by *Burkholderia pseudomallei*, a gram-negative bacillus classified as a Category B priority agent.

Platform: Dataset has 2 gpl platform. Sentrix Human-6 v2 Expression BeadChip and Illumina HumanHT-12 V3.0 expression beadchip

Study design:

1. Melioidosis patients (n=32).
2. Sepsis patients caused by other pathogens (n=31).
3. Uninfected control individuals (n=29).

Because we are investigating diagnostic genes for all kinds of sepsis we can Compare transcriptional profiles of sepsis patients (melioidosis and others) against uninfected controls to identify broader sepsis-related gene signatures.

1. Sepsis_limma_6947.csv -> sepsis patients 29 and healthy 5. (rows= 34, columns=91).
2. Sepsis_limma_6106.csv -> sepsis patients and healthy. (rows= 51, columns=69)

GPL6106:

In metadata they claimed the dataset had been normalized but it seems not truly normalized, so we used limma package for normalization between gpl and then filtered genes of interest and merged to pheno data to filter the sepsis patients and healthy controls among other patients with diabetes and control/recovery ones. (rows = 34 , columns = 90).

Present genes : 55
Missing genes : 0 ELANE as "ELA2", CXCL8 as “IL8”

Random Forest Model

Here is the result for sepsis patients and healthy ones in GPL6947. Random forest applied to the database 100 times by randomly splitting the dataset. (repeated_splits_metricsGSE13015-6947.csv)

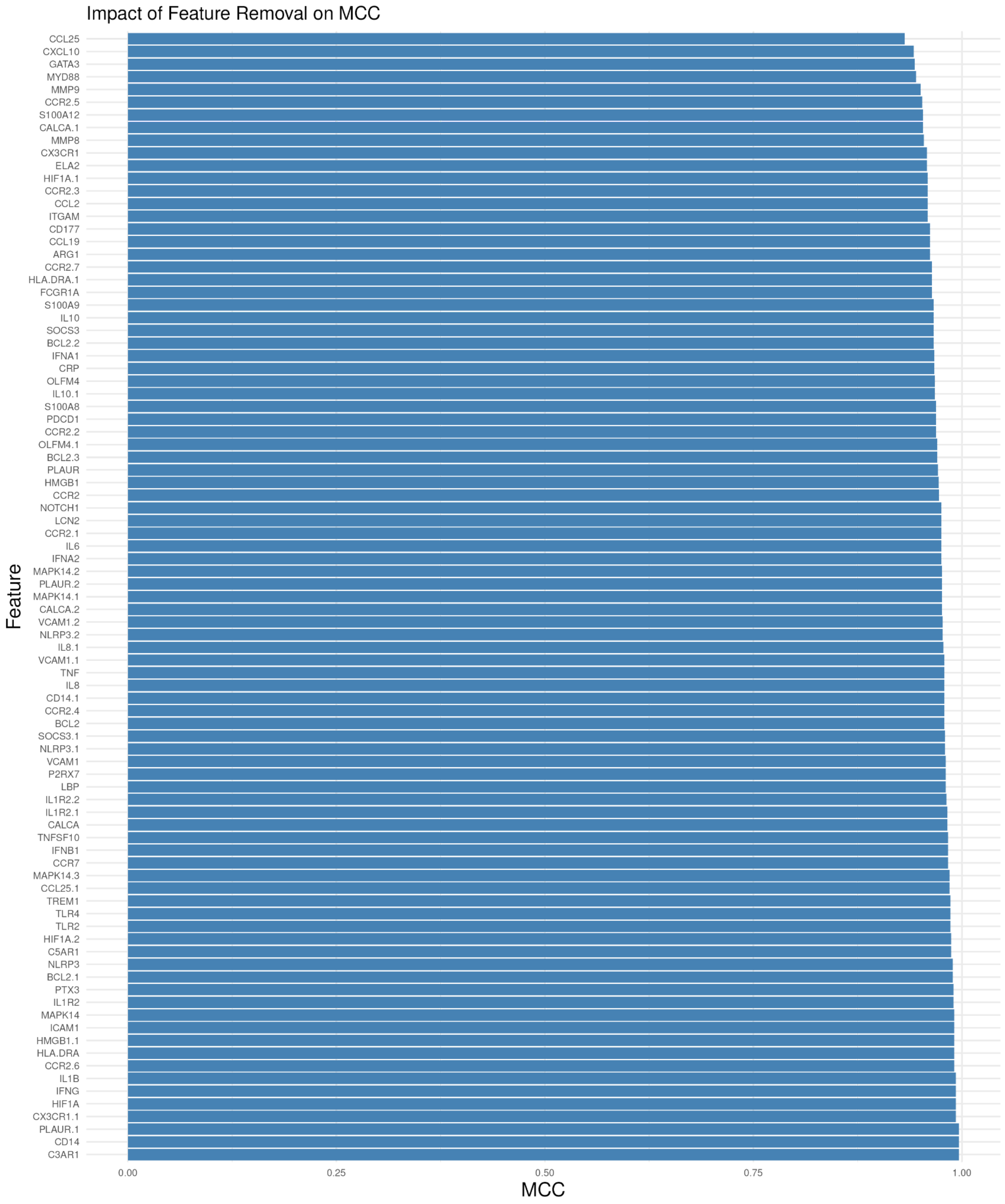
Average Metrics

The average of 100 times repeated for each metric is calculated . (average_metricsGSE13015-6947.csv)

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.949770357236347	0.963055555555556	0.97	0.958	0.97	0.97	0.943333333333333

Feature removal results:

Each feature has been removed from the dataset and all metrics values are calculated to find out the most important genes in true prediction. (feature_removal_resultsGSE13015-6947.csv)



1. Genes with High Impact (MCC drops significantly):(**MYD88, CXCL10, , MMP9, CCL25, GATA3**)

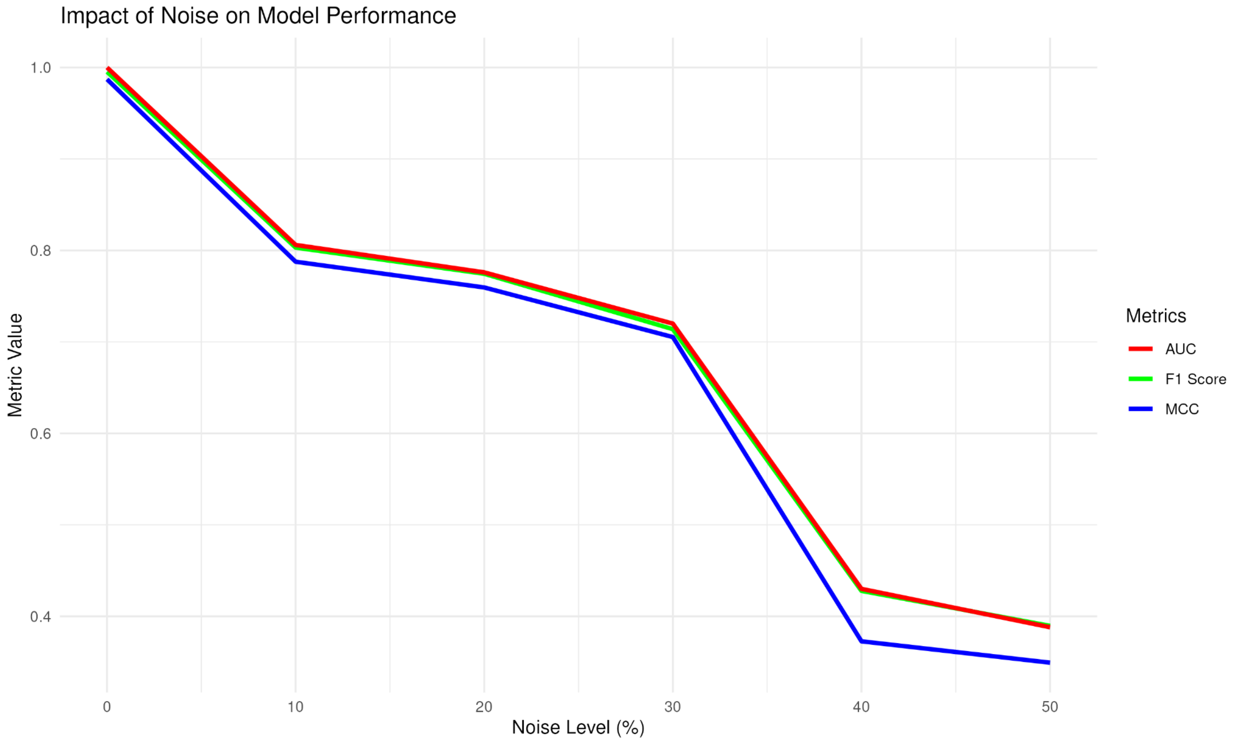
- Removing these genes causes a decline in MCC, indicating they are critical for the model's classification performance.
- These are likely the most informative genes for predicting the outcome (e.g., sepsis).

2. Genes with Low Impact (Minimal change in MCC):(**C3AR1, CD14, PLAUR.1,CX3CR1.1, HIF1A, IFNG)**

- Removing these genes has little to no effect on MCC, suggesting they are less important for the model’s classification task.

Sanity Check:

This plot represents the **sanity check** where different metrics are tracked as noise levels increase, with a dataset balanced using SMOTE (Synthetic Minority Oversampling Technique).(sanity_check_resultsGSE13015-6947.csv)



Initial Stability: At lower noise levels (near 0 on the X-axis), all metrics remain close to their maximum (around 1.0). This suggests the model performs well with minimal noise.

Gradual Decline: As noise increases, the performance starts to decline gradually. The green and red lines maintain slightly better performance than the blue line in the mid-range of noise levels.

Steep Drop-Off: Beyond a certain noise threshold (near 30-40 on the X-axis), performance drops significantly for all lines, showing that the models cannot handle high levels of noise effectively.

Final Performance: At the highest noise level (X=50), performance stabilizes at lower values (~0.4), indicating the models lose predictive power in this condition.

The main problem about the dataset is the target levels are imbalance, only 5 healthy controls and 29 Sepsis patients. After splitting data to train and test , the test data only contains 1 healthy and 5 sepsis.
I tried to do resampling by ROSE and SMOTE but it didn't work properly.

Also for **GPL6106** the dataset only includes 3 healthy controls and 48 sepsis patients. I couldn't use the resampling because there was a lot of error for a few healthy ones that it couldn't handle.

```
ANN: ERROR-------> Requesting more near neighbors than data points <-------------ERROR
```

I lowered the k from 5 even to 2 but still didnt work and it skipped the iterations.

```
Iteration 32 skipped due to class imbalance in test set or predictions.
Iteration 33 skipped due to class imbalance in test set or predictions.
```