Dataset description: GSE69063

The dataset **GSE69063** comprises gene expression profiles obtained from peripheral blood samples of patients experiencing **anaphylaxis**, **trauma**, **sepsis**, and healthy controls. The samples were collected as part of a study designed to investigate the gene expression patterns associated with these four disease phenotypes.

For this study, samples specifically related to **sepsis patients** and Healthy controls were extracted to identify gene expression changes .For the analysis, the dataset was separated based on the timeline (**T0**, **T1**, and **T2**) to assess temporal changes in gene expression.

The samples were collected at three different time points:

1. **T0**: At the time of emergency department (ED) arrival.

   (T0_sepsis_labeledGSE69063.csv) -> healthy control 33 and sepsis 19 samples.

2. **T1**: One hour after ED arrival.

   (T1_sepsis_labeledGSE69063.csv) -> healthy control 33 and sepsis 20 samples.

3. **T2**: Three hours after ED arrival.

   (T2_sepsis_labeledGSE69063.csv) -> healthy control 33 and sepsis 18 samples.

Presented-genes: 53
Missing-genes : FCGR1A, IFNA1

First we analyzed the T2 timeline which is 3 hours after a patient's ED arrival. (T2_sepsis_labeledGSE69063.csv).

Dataset for T2 by random forest target column as factor:

Random Forest Model:

Then a random forest model applied to the dataset by considering the target label as a factor.(RF-T2factorGSE69063code.R)
100 random splits are done. The result is saved in a csv file(repeated_splits_metrics.csv).

Average Model Metrics:(average_metrics.csv)

The overall performance metrics reflect a high-performing model:

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.9435079	0.9510000	1.0000000	0.9233333	1.0000000	1.0000000	0.9682143

MCC: 0.944 — Strong agreement between predictions and actual outcomes, even for imbalanced datasets.

F1: 0.951 — Excellent balance between precision and recall, showing high-quality predictions.

AUC: 1.000 — Perfect ability to distinguish between classes.

TPR (Sensitivity): 0.923 — High accuracy in identifying true positives.

TNR (Specificity): 1.000 — Perfect accuracy in identifying true negatives.

PPV (Precision for Positives): 1.000 — All positive predictions are correct.

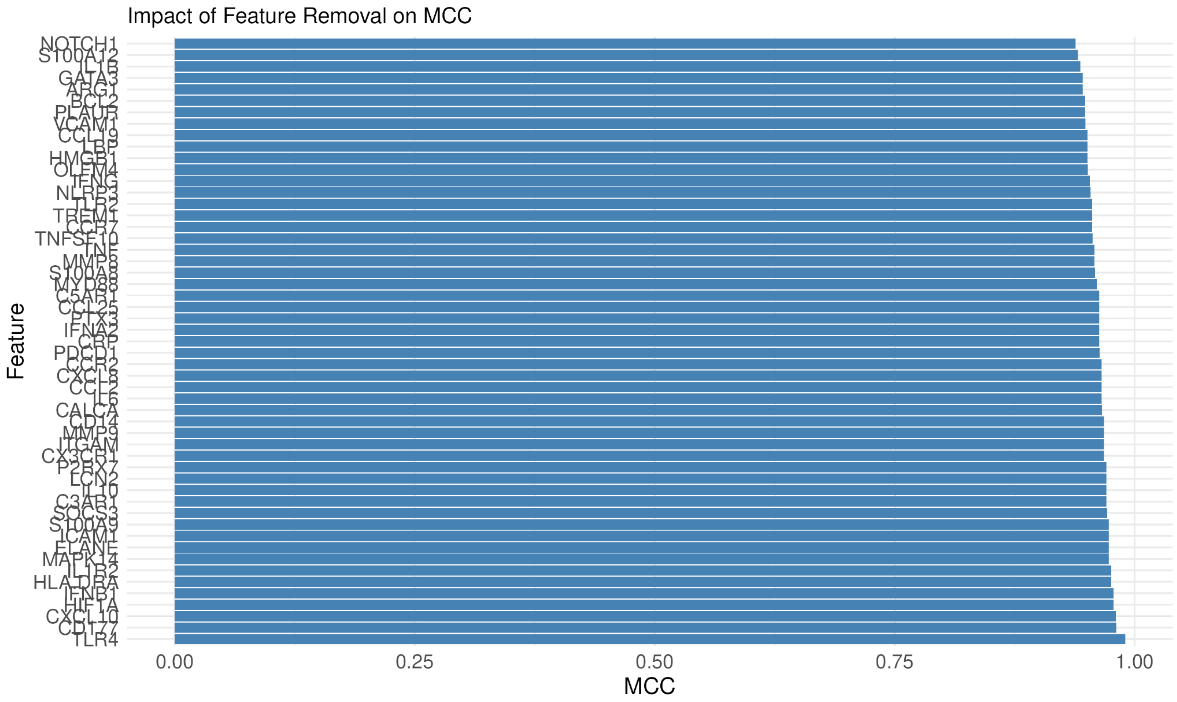
NPV (Precision for Negatives): 0.968 — Strong ability to predict negatives accurately.

The model shows near-perfect performance.

Feature Removal Analysis:

This feature removal plot provides insights into the importance of individual features (genes) in contributing to the model's performance. The x-axis represents the MCC (Matthews Correlation Coefficient), and the y-axis lists the features (genes).

The results of removing individual features and observing the impact on metrics highlight the importance of different genes for predictions.The results has been saved in (feature_removal_results.csv) and here is the plot for impact for feature removal on MCC:



Most Important Genes (TOP of the Plot):

These genes cause the largest drop in performance (MCC) when removed, indicating they are critical for the model's predictions.(**NOTCH1, S100A12, IL1B, ARG1, GATA3**).

Least Important Genes (Bottom of the Plot):

These genes cause the smallest drop in performance, meaning their contribution is minimal or redundant in the model.(**CD14, TLR4, IFNB1, CX3CR1, CRP**)

Moderately Important Genes (Middle of the Plot):

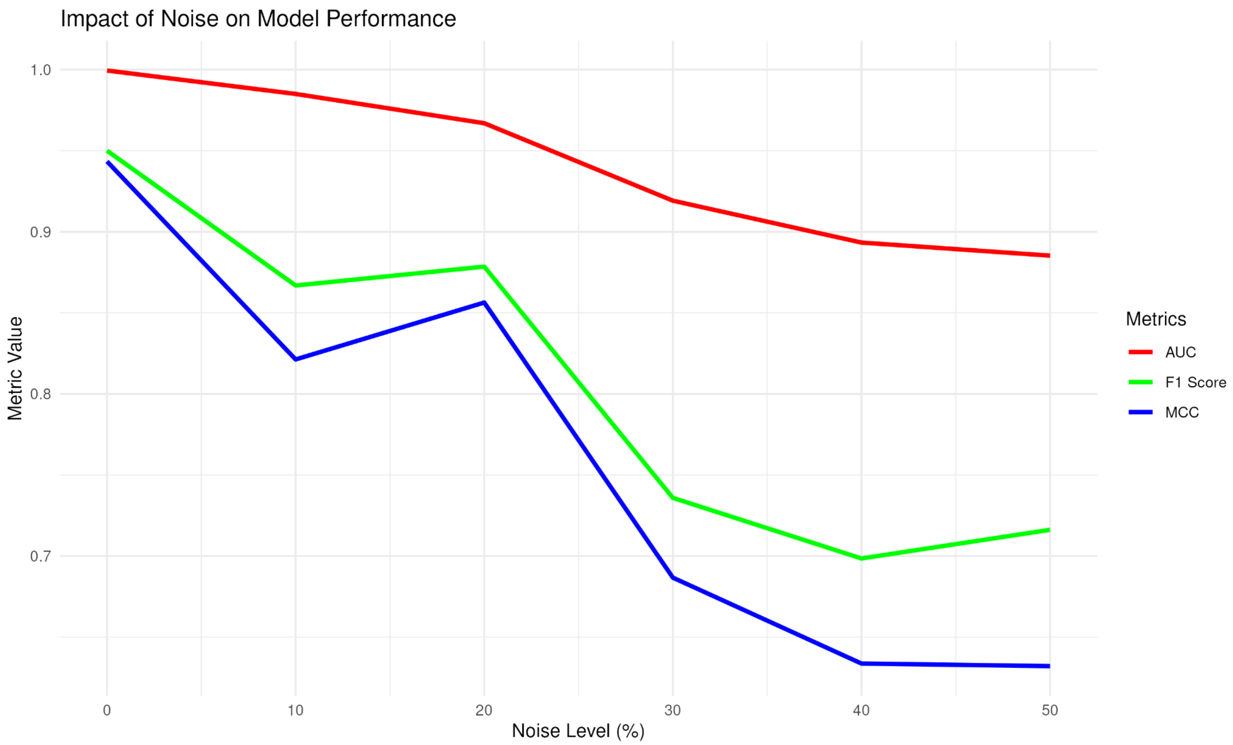
These genes cause a moderate drop in performance, suggesting they are important but somewhat redundant or less critical than the top genes.

1. **CALCA**
2. **MAPK14**
3. **LCN2**
4. **IL10**
5. **PTX3**

Noise Robustness Test:

The model's performance under increasing levels of noise (as seen in the plot) shows:

(sanity_check_results.csv)



- Noise Level 0 (Baseline):
  - Perfect or near-perfect metrics (e.g., AUC = 1.0, MCC = 0.944) with clean data.
- Moderate Noise (Noise Levels 10–20):
  - Metrics degrade slightly but remain strong, reflecting robustness to minor data disruptions.
- High Noise (Noise Levels 30–50):
  - Significant drops in metrics:
    - MCC drops sharply to ~0.700 at Noise Level 30 and further to ~0.600 at Noise Level 50.
    - AUC remains relatively robust but shows gradual decline, reflecting resilience in distinguishing between classes despite noise.
    - F1, TPR, and TNR metrics show sensitivity to high noise, indicating the model struggles more to classify accurately under noisy conditions.

Dataset for T2 by random forest target column as number(0 and 1):

Random Forest Model:

Then a random forest model applied to the dataset by considering the target label as numbers. (RF-num_labelGSE69063code.R)

100 random splits are done. The result is saved in a csv file(repeated_splits_metrics.csv).

Average Model Metrics:(average_metrics.csv)

The overall performance metrics reflect a high-performing model:

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.930387935433589	0.941955988455988	0.98	0.917166666666667	0.98	0.98	0.945896825396825

MCC: 0.93 — has also Strong agreement between predictions and actual outcomes, even for imbalanced datasets.

F1: 0.94 — Excellent balance between precision and recall, showing high-quality predictions.

AUC: 0.98 — Perfect ability to distinguish between classes.

TPR (Sensitivity): 0.91 — High accuracy in identifying true positives.

TNR (Specificity): 0.98 — Perfect accuracy in identifying true negatives.

PPV (Precision for Positives): 0.98 — All positive predictions are correct.

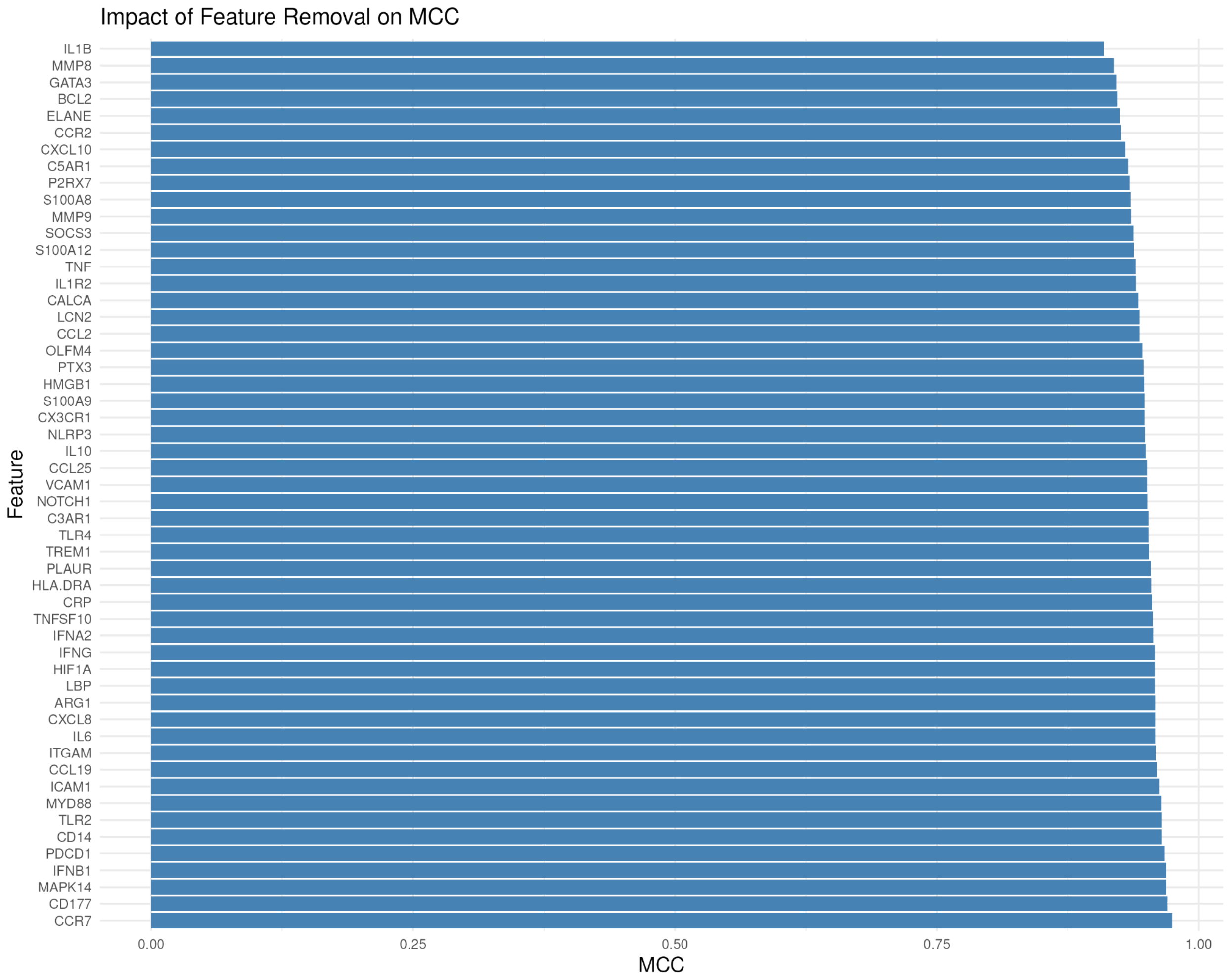
NPV (Precision for Negatives): 0.94 — Strong ability to predict negatives accurately.

The model shows near-perfect performance. But it shows slight decrease(unnoticeable) in the results but its still perfect.

Feature Removal Analysis:

This feature removal plot provides insights into the importance of individual features (genes) in contributing to the model's performance. The x-axis represents the MCC (Matthews Correlation Coefficient), and the y-axis lists the features (genes).

The results of removing individual features and observing the impact on metrics highlight the importance of different genes for predictions.The results has been saved in (feature_removal_results.csv) and here is the plot for impact for feature removal on MCC:



Most Important Genes (TOP of the Plot):

These genes cause the largest drop in performance (MCC) when removed, indicating they are critical for the model's predictions.(**MMP8, ELANE, IL1B, BCL2, GATA3**).

Least Important Genes (Bottom of the Plot):

These genes cause the smallest drop in performance, meaning their contribution is minimal or redundant in the model.(**CD14, CD177, MAPK14, IFNB1,PCCD1, CCR7**)

Moderately Important Genes (Middle of the Plot):

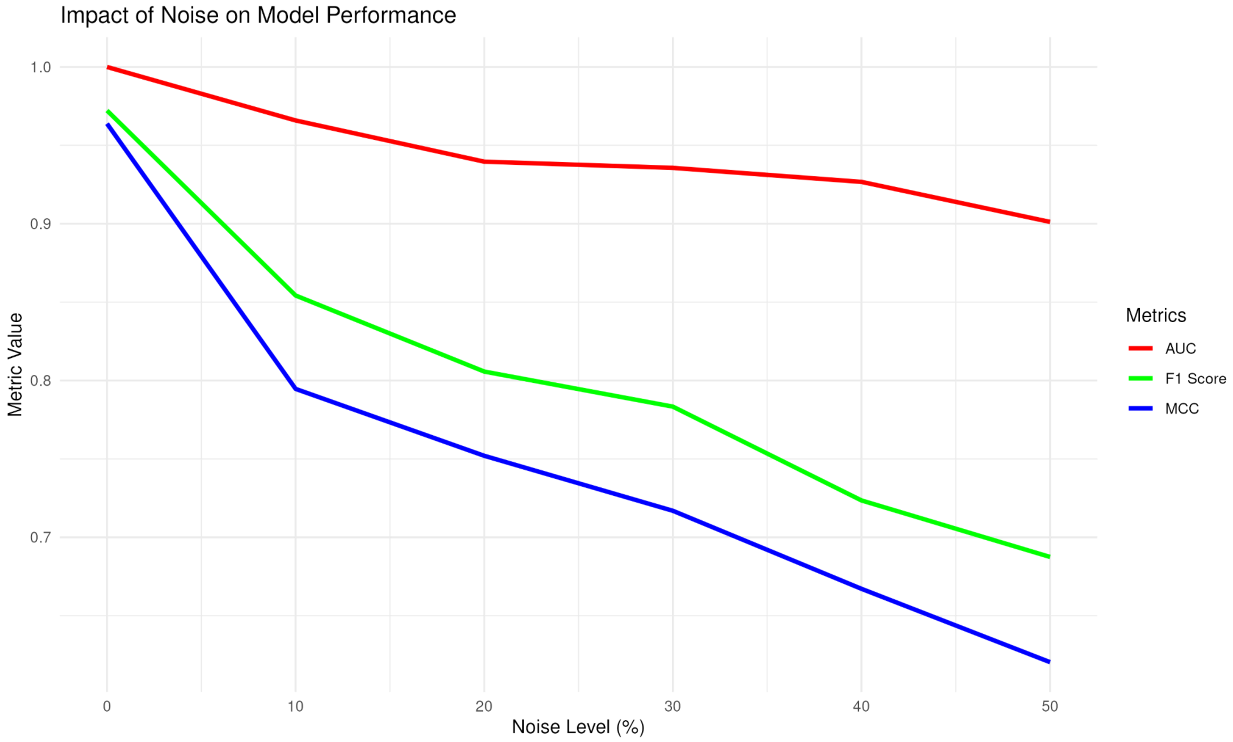
These genes cause a moderate drop in performance, suggesting they are important but somewhat redundant or less critical than the top genes.

6. **LCN2**
7. **VCAM1**
8. **CRP**
9. **CCL25**
10. **NOTCH1**

Noise Robustness Test:

The model's performance under increasing levels of noise (as seen in the plot) shows:

(sanity_check_results.csv)



- Noise Level 0 (Baseline):
    - Perfect or near-perfect metrics (e.g., AUC = 1.0, MCC = 0.90) with clean data.
- Moderate Noise (Noise Levels 10–20):
    - Metrics degrade slightly but remain strong, reflecting robustness to minor data disruptions.
- High Noise (Noise Levels 30–50):
    - Significant drops in metrics:
        - MCC drops sharply to ~0.700 at Noise Level 30 and further to ~0.550 at Noise Level 50.
        - AUC remains relatively robust but shows gradual decline, reflecting resilience in distinguishing between classes despite noise.
        - F1, TPR, and TNR metrics show sensitivity to high noise, indicating the model struggles more to classify accurately under noisy conditions.