Dataset: **GSE54514**

This dataset, GSE54514, is a gene expression profiling study focused on whole blood samples collected from critically ill patients. The primary objective of the study is to assess the potential of gene expression profiling as a diagnostic and monitoring tool for immune dysfunction in septic patients.

Platform: **Expression profiling by array** using the **Illumina HT-12 gene expression microarray** platform.

Study Population:

The dataset includes three groups:

Sepsis survivors (n = 26)

Sepsis non-survivors (n = 9)

Healthy controls (n = 18)

The dataset also have information about survivor and deceased patients so we extract 2 kinds of dataset from the whole data:(GSE54514code.R)

1. Diagnosis dataset which includes 127 sepsis patients and 36 healthy ones. (sepsis_diagnosis_dataGSE54514.csv). 163 rows and 76 columns.

2. Prognostic dataset which includes 31 Nonsurvivor and 96 Survivor patients. (sepsis_prognosis_dataGSE54514.csv). 127 rows and 76 columns.

The number of samples in the dataset description is different from the actual number of samples in the dataset. I think they may include samples from patients within 5 days but they are not truly labeled in metadata related to patients. So I think the dataset needs more investigation.

The dataset was already normalized , here are the results for a random forest model on sepsis diagnosis dataset. The dataset is imbalanced in the number of sepsis patients and healthy control so we used the ROSE for resampling but it didn't work properly and we used SMOTE to solve this issue. (RF-factor-smote-GSE54514code.R).

Random Forest Model

Here is the result for patients from sepsis patients and healthy ones . Random forest applied to the database 100 times by randomly splitting the dataset.
(repeated_splits_metrics_with_SMOTE54514.csv)

Average Metrics

The average of 100 times repeated for each metric is calculated .
(average_metrics_with_SMOTE54514.csv)

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.5702904224876512	0.9158445988487672	0.9126571428571432	0.952	0.5457142857142862	0.8842103973487782	0.7851904761904762

MCC: Interpretation: A score of 0.57 suggests moderate predictive power. This is acceptable but indicates room for improvement.

F1 score: A high F1 score of **0.916** indicates good balance between precision and recall, meaning the model effectively identifies true positives with relatively low false positives.

AUC: A score of 0.913 shows that model has excellent discriminatory ability, effectively separating positive and negative classes.

TPR: A TPR of 0.952 indicates that the model is very good at identifying positive cases, missing only 4.8% of them.

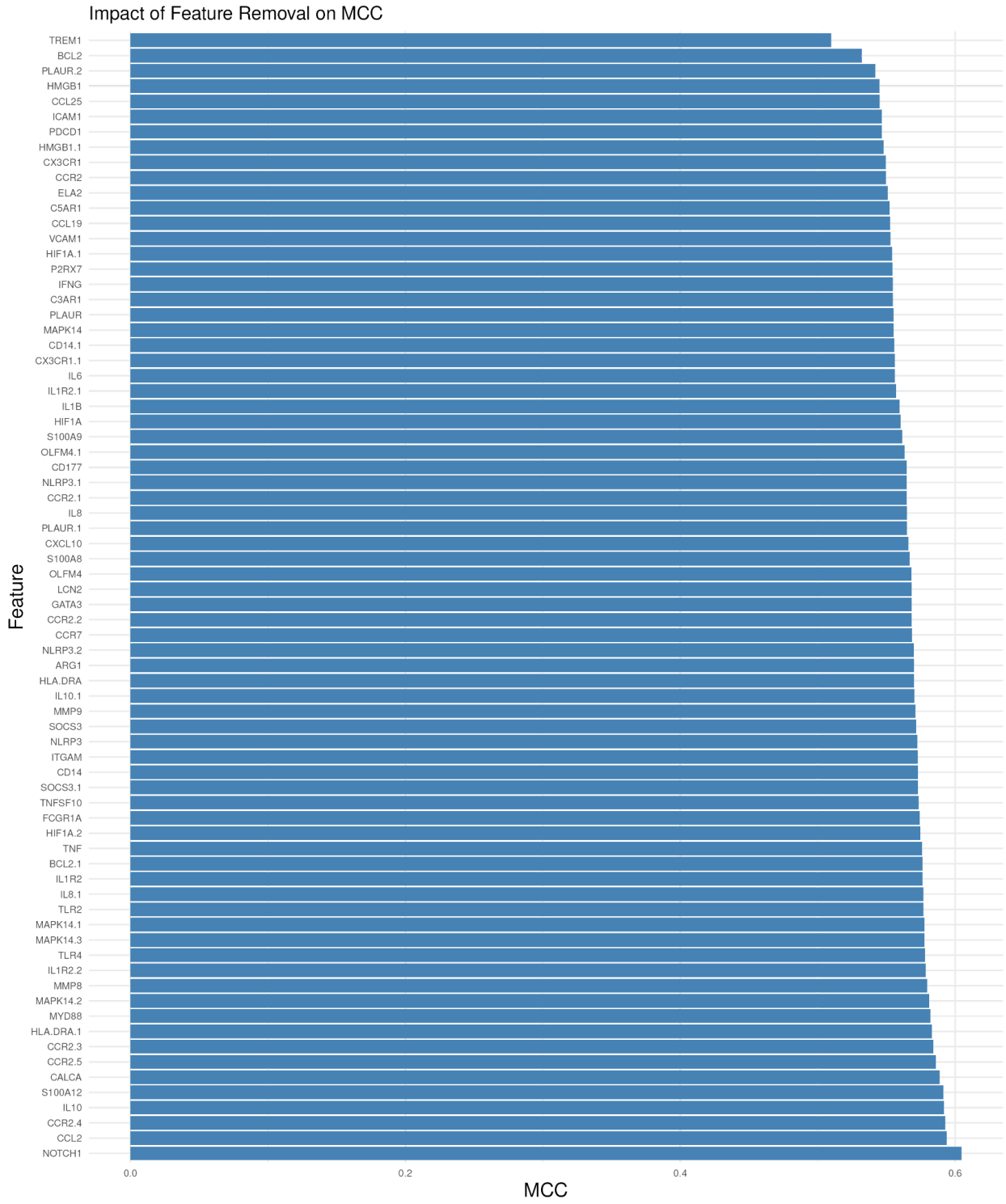
TNR: A TNR of 0.546 suggests the model struggles to identify negative cases correctly, meaning it frequently misclassified negatives as positives.

PPV: A PPV of 0.884 shows that the majority of positive predictions are correct, with some false positives present.

NVP: An NPV of 0.785 indicates that a significant proportion of negative predictions are correct, though some false negatives exist.

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_results_with_SMOTE54514.csv)



1. Genes with High Impact (MCC drops significantly):(**TREM1, CBCL2, PLAUR.2, HMGB1, CCL25, ICAM1**)

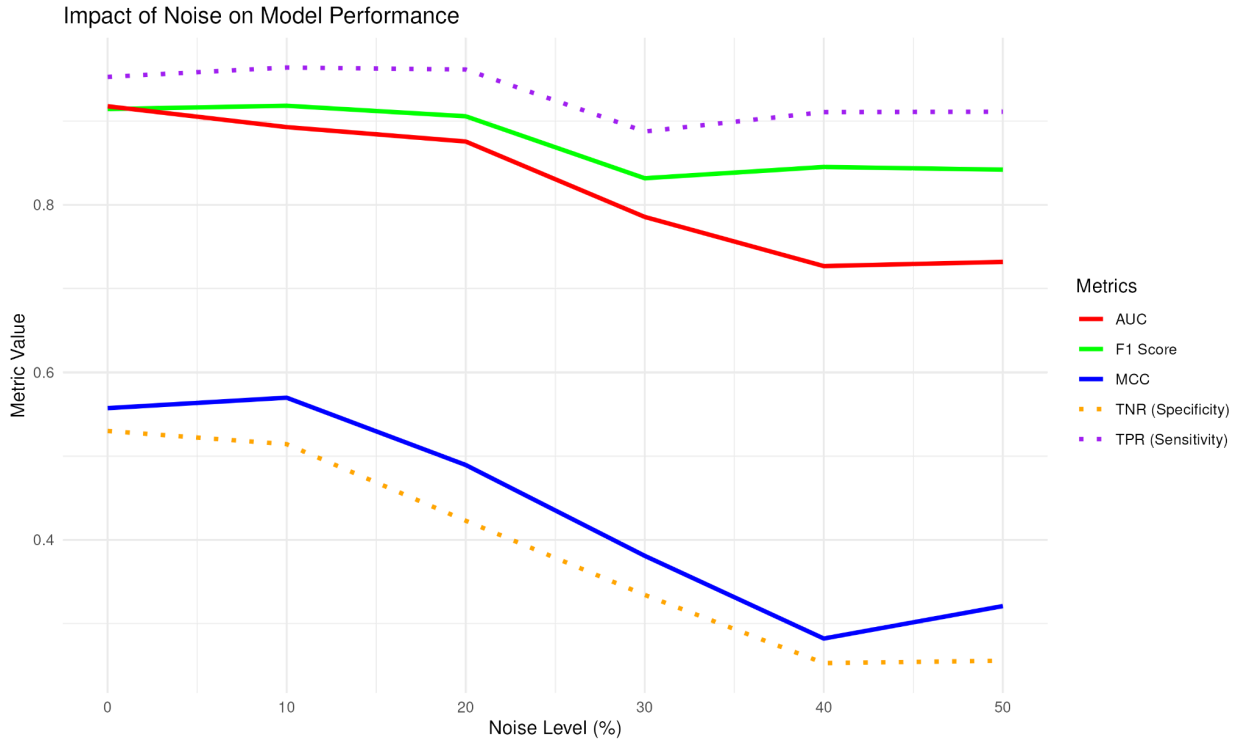
- Removing these genes causes a sharp decline in MCC, indicating they are critical for the model's classification performance.
- These are likely the most informative genes for predicting the outcome (e.g., sepsis).

2. Genes with Low Impact (Minimal change in MCC):(**NOTCH1, CCL2, CCR2.4, IL10, S100A12, CALCA)**

- 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_results_with_SMOTE54514.csv)



Impact of Noise on Metrics:

- **AUC and F1 Score** are more robust to noise compared to MCC and TNR, which degrade significantly. This suggests the model's ability to rank predictions (AUC) and balance precision/recall (F1 Score) is less affected by noise.

- **MCC and TNR** drop significantly, indicating that the model's ability to balance between positive and negative cases deteriorates under noisy conditions, especially for identifying negatives (controls).

Positive Case Handling:

- The **TPR** remains stable, showing the model performs well in identifying positive cases (e.g., sepsis) even with noise.

Negative Case Handling:

- The sharp decline in **TNR** indicates the model struggles with identifying negatives (e.g., healthy controls or non-sepsis cases) as noise increases. This could be due to overlaps in feature distributions caused by noise.
