Dataset: GSE137342

This dataset is part of a large-scale clinical transcriptomic analysis focused on sepsis patients in an Indian cohort. The study investigates blood gene expression in patients suffering from severe sepsis and septic shock admitted to the ICU. The goal of this research was to understand the transcriptional changes associated with sepsis and septic shock.

By utilizing transcriptional module scores, the study quantifies biological processes related to immune response, inflammation, and sepsis progression. The findings indicate that genes related to immune response were more suppressed, while inflammation-related genes were more active in sepsis patients. These insights could be valuable for developing diagnostic biomarkers and prognostic tools for sepsis management.

Overall Design:

- **Patients and Controls:**
  - The study included **27 sepsis patients** and **12 healthy controls**.
  - The healthy controls were **age- and gender-matched** but had no inflammatory diseases.
- **Time Points for Sepsis Patients:**
  - **Day 1 (D1):** Blood samples were collected at the time of sepsis diagnosis.
  - **Day 2 (D2):** A second blood sample was collected **24 hours after diagnosis** to observe changes in gene expression over time.
- **Healthy Control Samples:**
  - A **single blood sample** was collected from each healthy individual.

The dataset has 2 GPLs which GPL 16686 does not include the healthy control ones, so we just investigated the GPL10558.

For sepsis patients they have mixed all sepsis, severe sepsis and septic shock sampling in Day1 and Day2 in the dataset. So we just separate each dataset for each day:

Day1 ->

1. **sepsis_allD1_GSE137342_10558.csv** that includes all sepsis , severe sepsis and septic shock patients with healthy controls.

2. **sepsis_onlyD1_GSE137342_10558.csv** that includes only sepsis and severe sepsis and healthy controls.

3. **septic_shockD1_GSE137342_10558.csv** that includes only septic shock patients and healthy controls.

Day 2 ->

1. **sepsis_allD2_GSE137342_10558.csv** that includes all sepsis , severe sepsis and septic shock patients with healthy controls.

2. **sepsis_onlyD2_GSE137342_10558.csv** that includes only sepsis and severe sepsis and healthy controls. 15 sepsis and 12 healthy samples. (27 rows and 86 columns)

3. **septic_shockD2_GSE137342_10558.csv** that includes only septic shock patients and healthy controls.

Present genes: 53

Missing genes: 2 (PTX3, MMP8)

---

sepsis_onlyD2_GSE137342_10558:

The dataset was already normalized. First the random forest applied to samples from day 2 of only sepsis patients and healthy ones.(sepsis_onlyD2_GSE137342_10558.csv). Because the dataset was almost balanced we didn't use the SMOTE resampling. (Rf-sep-onlyD2-GSE137342-10558code.R)

Random Forest Model:

Random forest applied to the dataset 100 times by randomly splitting the dataset to train and test data by 0.8 to predict sepsis patients based on their gene expression data. (repeated_splits_metrics_D2_seponly.csv)

Average Metrics

The average of 100 times repeated for each metric (MCC, F1 score, AUC, TPR, TNR,...) is calculated . (average_metrics.csv)

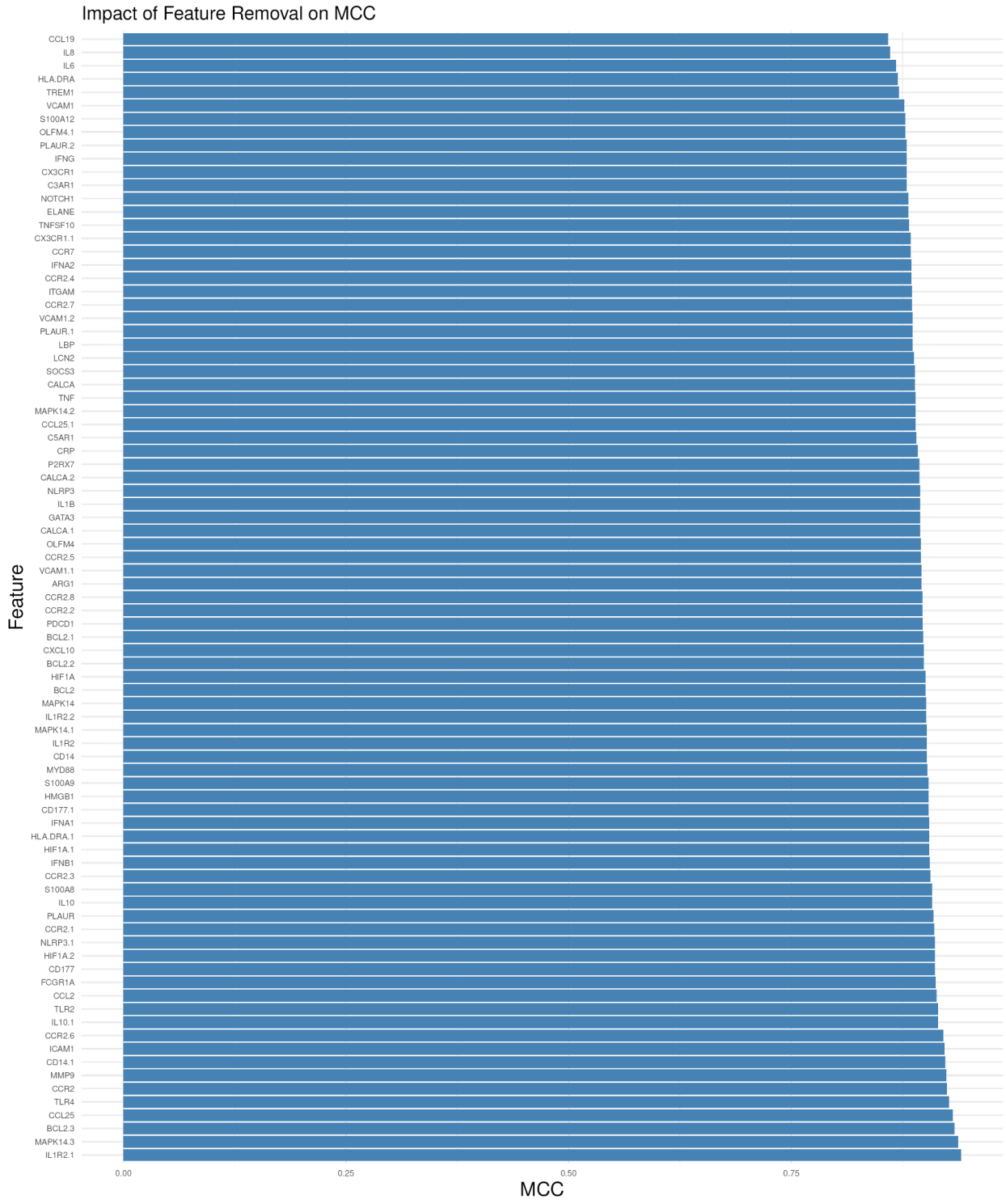
MCC	F1	AUC	TPR	TNR	PPV	NPV
0.924831632475944	0.952	1	0.923333333333333	1	1	0.926666666666667

Overall Assessment

- **MCC (0.92)**: 0.9248 indicates strong agreement between predicted and actual labels, meaning the model has high predictive power.
- **F1 Score (0.95)**: 0.952 means the model maintains an excellent balance between precision and recall, meaning it correctly predicts most sepsis cases without many false positives.
- **AUC (1)**: Perfect class separation.
- **TPR (0.92)** and **TNR (1)**: High sensitivity and specificity.
- **PPV (1)** and **NPV (0.92)**: Accurate predictions for both classes. 0.9267 (92.67%) means that when the model predicts a patient is healthy, it is correct 92.67% of the time.

Feature removal results:
Each feature has been removed from the dataset and all metrics values are calculated in each iteration to find out the most important genes that impact the MCC value by its removal from the dataset. (feature_removal_results.csv)

The plot shows us MCC calculated after removing each feature(gene) from the dataset. (Impact of feature removal on MCC)



1. Genes with High Impact (MCC drops significantly):(**CCL19, IL8 , IL6, TREM1, VCAM1, HLA-DRA**)

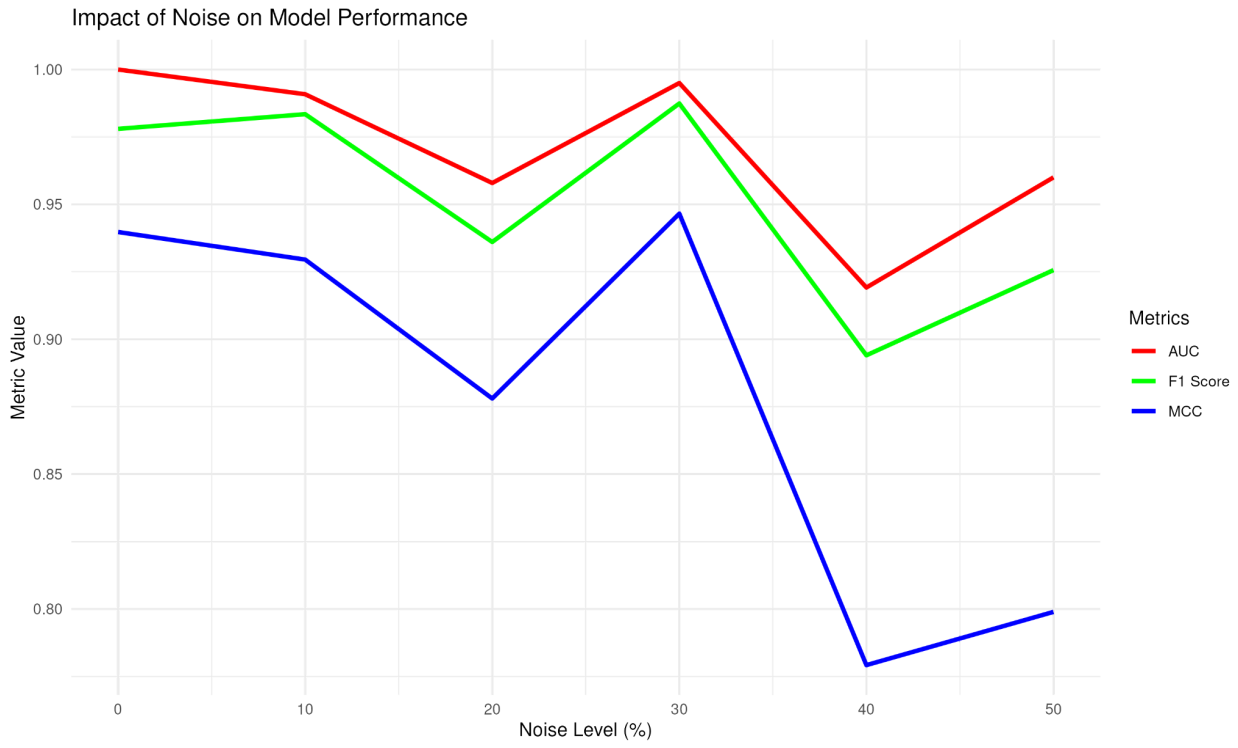
- Removing these genes causes a sharp decline in MCC, indicating they are critical for the model's classification performance.

- These are likely the most informative genes for predicting the outcome (e.g., sepsis).

2. Genes with Low Impact (Minimal change in MCC):(**MAPK14.3, BCL2.3, CCL25, IL1R2.1, TLR4, CCR2)**
- 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.(sanity_check_results_D2_seponly.csv)



Impact of Noise on Metrics:

The model performs well under low noise conditions but degrades significantly after 30% noise. The AUC remains relatively high, meaning the model still ranks sepsis vs. non-sepsis cases well. MCC suffers the most, indicating that noise impacts the overall reliability of predictions.

But all in all the dataset is small and this the number of each label in the dataset:

```
table(data$Label)


Healthy  Sepsis

     12      15

> table(test_data$Label)


Healthy  Sepsis

      2       3

> table(train_data$Label)


Healthy  Sepsis

     10      12
```

---

sepsis_allD2_GSE137342_10558

Then the random forest is again applied on the dataset that we prepared and includes both sepsis and septic shock patients and healthy controls.(sepsis_allD2_GSE137342_10558.csv)

(Rf-allD2-GSE137342-10558code.R)

Random Forest Model:

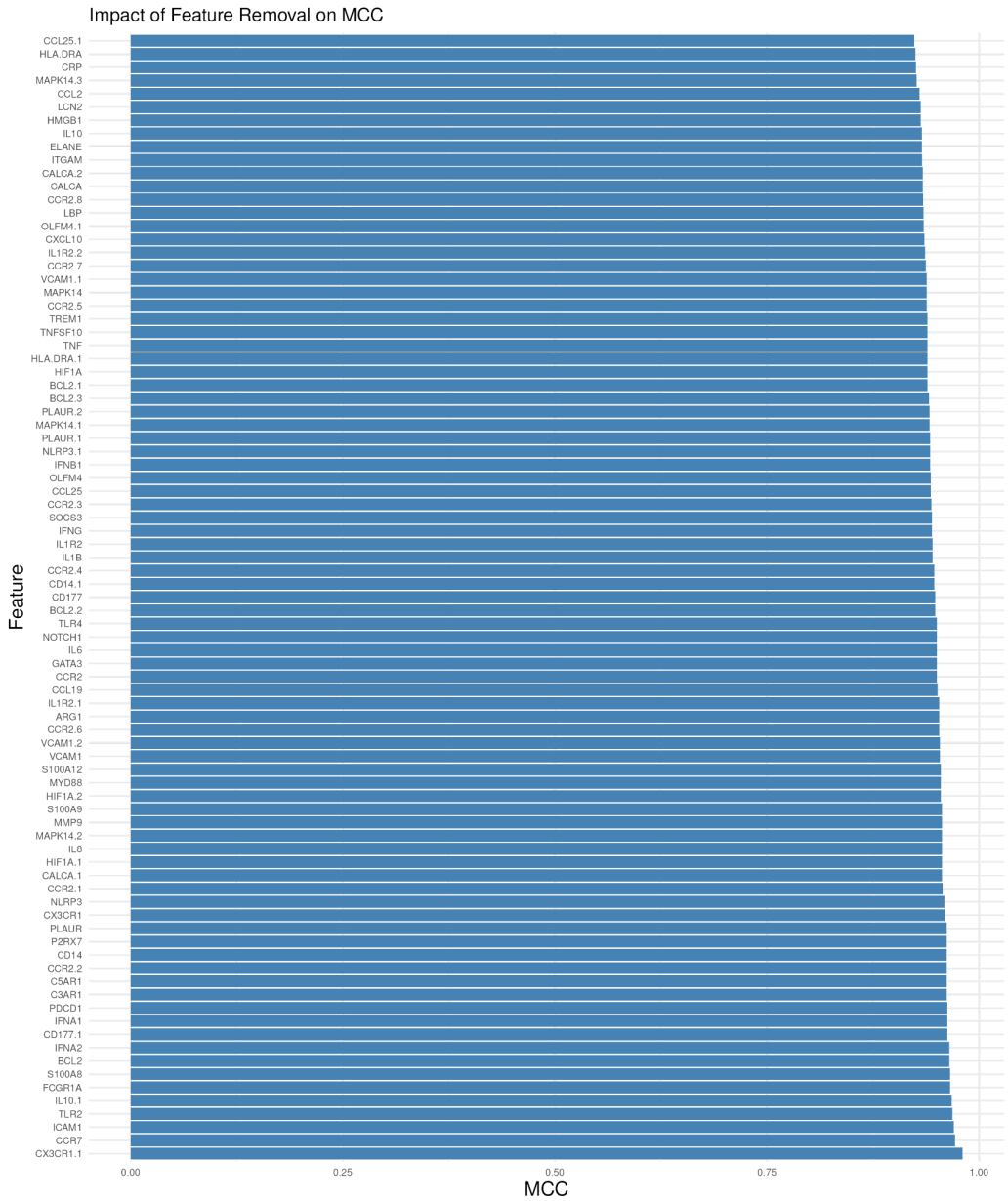
Random forest applied to the dataset 100 times by randomly splitting the dataset to train and test data by 0.8 to predict sepsis patients based on their gene expression data. (repeated_splits_metrics_allD2.csv)

Average Metrics
The average of 100 times repeated for each metric (MCC, F1 score, AUC, TPR, TNR,...) is calculated . (average_metrics_allD2.csv)

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.95985281 3742386	0.97952380 9523809	1	0.965	1	1	0.955

Feature removal results:

Each feature has been removed from the dataset and all metrics values are calculated in each iteration to find out the most important genes that impact the MCC value by its removal from the dataset. (feature_removal_results_allD2.csv)



3. Genes with High Impact (MCC drops significantly):(**CCL25.1, CRP , MAPK14.3, CCL2, LCN2, HLA-DRA**)

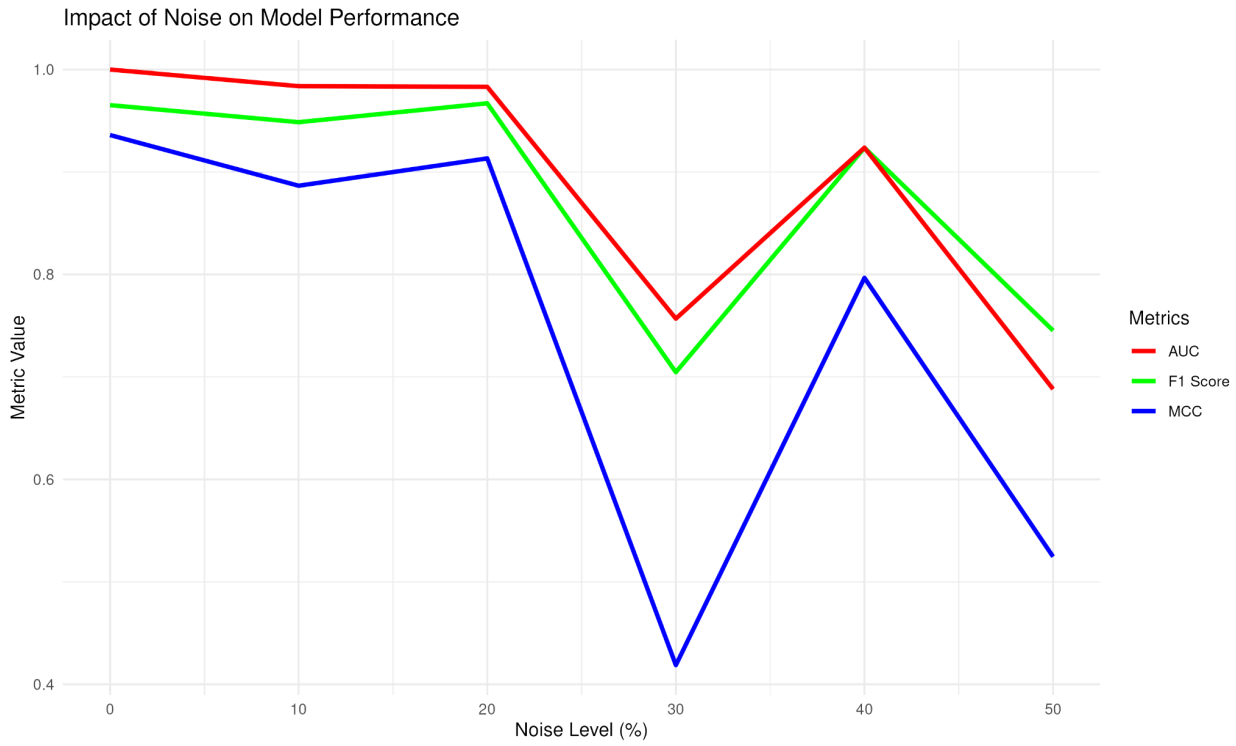
- Removing these genes causes a sharp decline in MCC, indicating they are critical for the model's classification performance.
- These are likely the most informative genes for predicting the outcome (e.g., sepsis).

4. Genes with Low Impact (Minimal change in MCC):(**CX3CR1.1, CCR7, ICAM1, TLR2, IL10.1, FCGR1A)**

- 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.(sanity_check_results_allD2.csv)



Impact of Noise on Metrics:

The model performs well under low noise conditions but degrades significantly after 30% noise. The AUC remains relatively high, meaning the model still ranks sepsis vs. non-sepsis cases well. MCC suffers the most, indicating that noise impacts the overall reliability of predictions.

Because the dataset was imbalanced in the number of sepsis and healthy ones , we performed a resampling technique to make a balance between our target labels.

```
table(data$Label)
```

```
Healthy  Sepsis
     12      22
```

At first we performed the ROSE resampling which did not result properly. The average metrics became
MCC : -0.8929899

F1 : 0.3181818

AUC : 0.99875

TPR : 0.095

TNR : 0

PPV : 0.1183333

NPV : 0

> table(test_data$Label) Healthy Sepsis 2 4

> table(train_data$Label) Sepsis Healthy 14 14

The results may be because ROSE uses a more general approach for generating new samples via a kernel density estimate or random perturbations, so it can potentially create more diverse synthetic examples but also sometimes less “structured” ones, and ROSE may oversample minority instances and undersample majority class in a broader manner, sometimes producing more “random” points, which can be good or bad depending on the structure of your data.

So we performed SMOTE resampling because For each minority class example, SMOTE finds its *k* nearest neighbors (in feature space) and randomly interpolates between them to create synthetic points. SMOTE focuses on minority points and tries to stretch or “fill in” the region around them. This can improve how the model learns the minority class boundary. Thus SMOTE can be a great alternative if we have numeric data and want to create more “realistic” synthetic minority points by interpolation.

sepsis_allD2_GSE137342_10558_SMOTE

Then the random forest is again applied on the dataset that we prepared and includes both sepsis and septic shock patients and healthy controls.(sepsis_allD2_GSE137342_10558.csv)

(Rf-smote-allD2-GSE137342-10558code.R)

Random Forest Model:

Random forest applied to the dataset 100 times by randomly splitting the dataset to train and test data by 0.8 to predict sepsis patients based on their gene expression data. (repeated_splits_metrics_smote_allD2.csv)

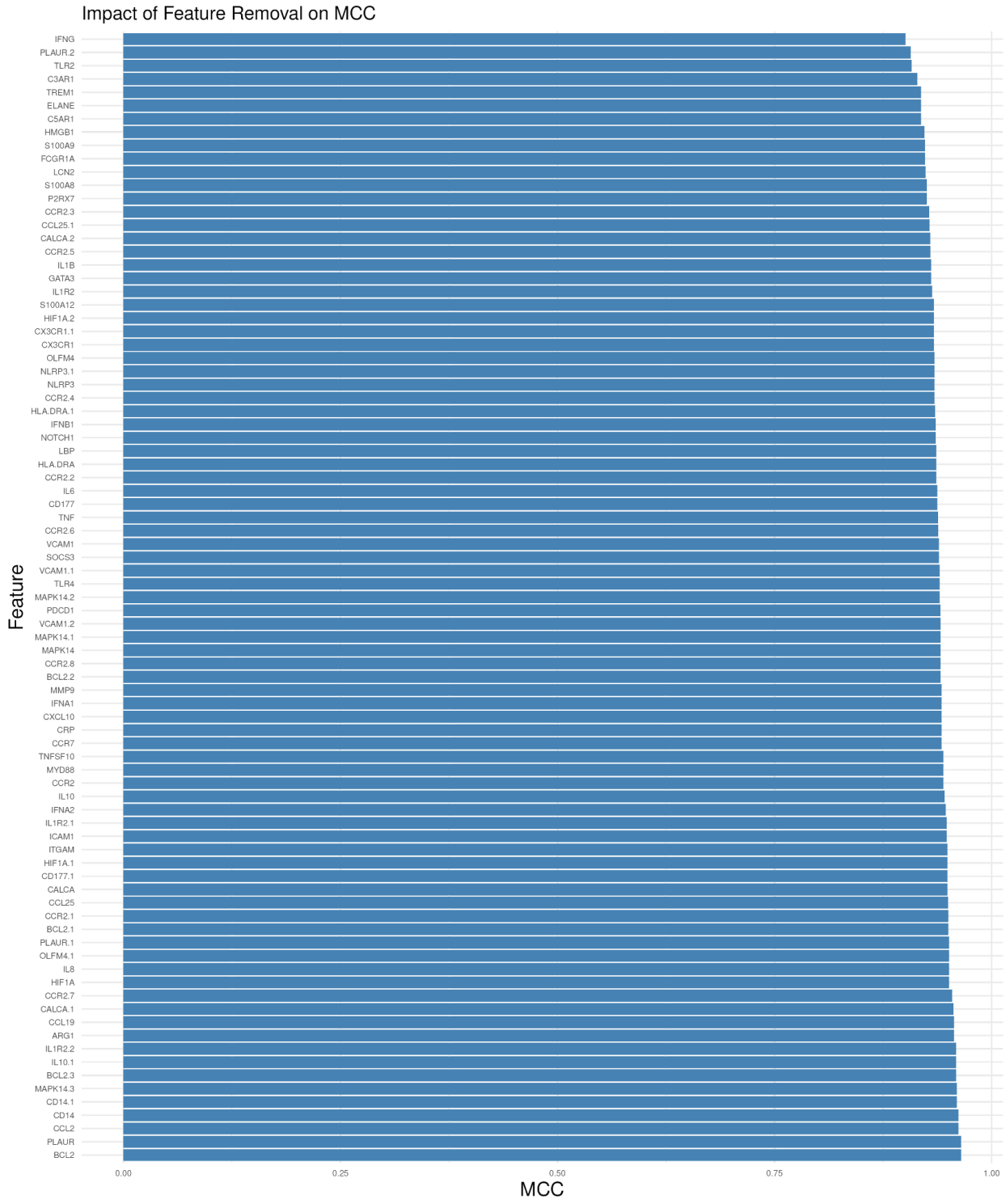
Average Metrics

The average of 100 times repeated for each metric (MCC, F1 score, AUC, TPR, TNR,...) is calculated . (average_metrics_smote_allD2.csv)

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.93142135 6237309	0.96476190 4761905	0.99875	0.94	1	1	0.92333333 3333333

Feature removal results:

Each feature has been removed from the dataset and all metrics values are calculated in each iteration to find out the most important genes that impact the MCC value by its removal from the dataset. (feature_removal_results_smote_allD2.csv)



5. Genes with High Impact (MCC drops significantly):(**IFNG, PLAUR.2 , TLR2, C3AR1, TREM1, ELANE**)

- 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).

6. Genes with Low Impact (Minimal change in MCC):(**BCL2, PLAUR, CCL2, CD14, CD14.1, MAPK14.3)**
- 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.(sanity_check_results_smote_allD2.csv)