Dataset: GSE100159

This dataset represents the development of a cost-effective and practical targeted transcriptional fingerprinting assay (TFA) for monitoring diverse immune responses in humans. The assay leverages a modular transcriptional repertoire framework based on variations in transcript abundance measured on a genome-wide scale in blood samples. The dataset comprises nearly 1000 human subjects across 16 immunologic conditions.

The assay is particularly aimed at detecting transcriptional variations that can inform immune responses in conditions such as sepsis, infections, and other immune-related disorders. It focuses on targeted profiling to enable clinically actionable insights.

Overall Design:

The dataset includes 47 total samples, divided as follows:(2 recovery from sepsis)

Sepsis group: 35 samples

Control group: 12 samples

This design allows for comparisons between sepsis patients and controls to identify transcriptional differences that could serve as biomarkers or indicators of sepsis.

Present genes : 55

Missing genes : 0 ELANE exist as “ELA2” and CXCL8 as “IL8”

Dataset had been already normalized then data related to only sepsis patients and control ones was extracted. (sepsis_dataGSE100159.csv) it includes (45 rows and 86 columns) because some genes have different probes. 12 sepsis and 33 controls.

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_metrics100159.csv)

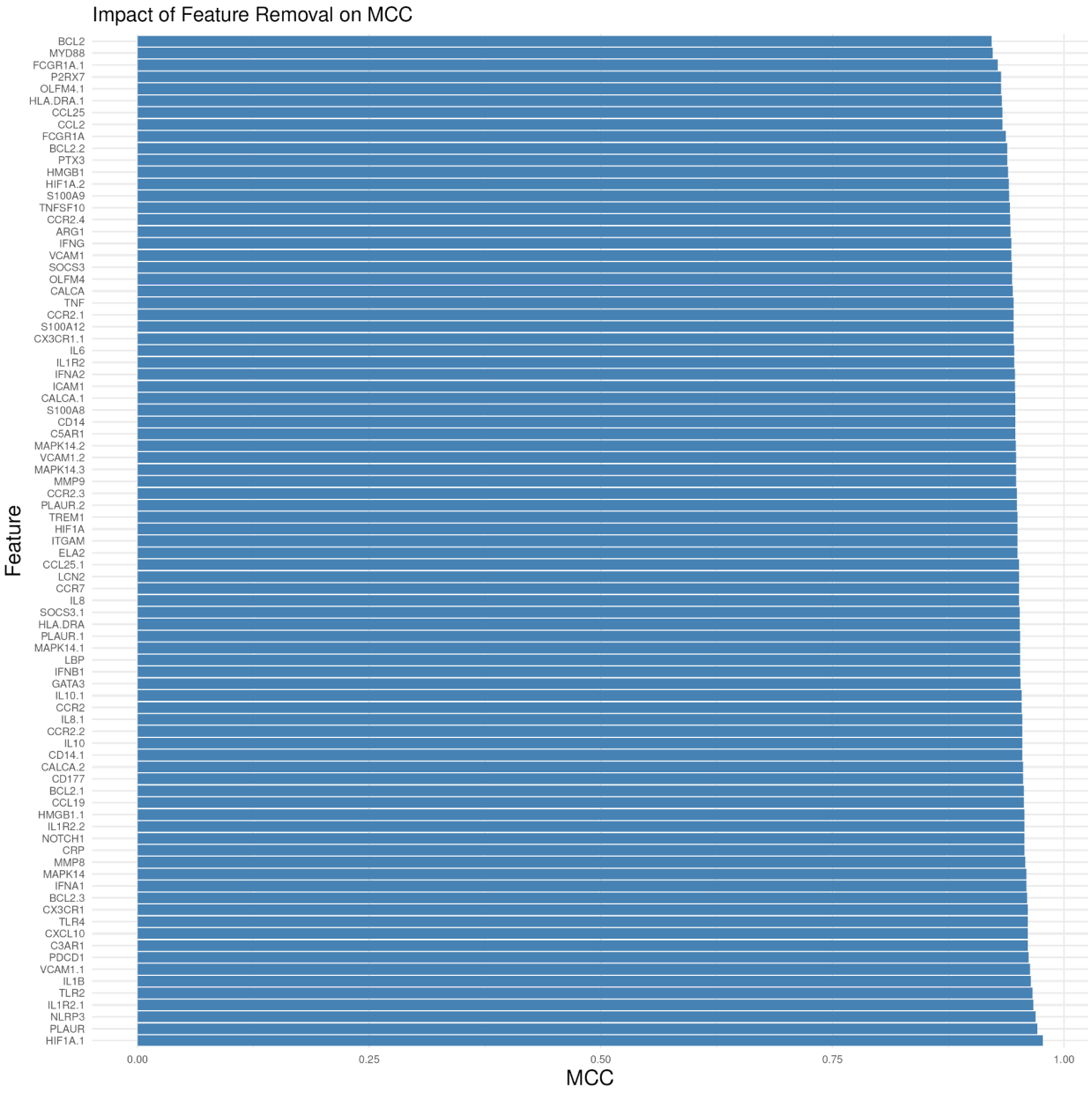
Average Metrics
The average of 100 times repeated for each metric (MCC, F1 score, AUC, TPR, TNR,...) is calculated . (average_metrics100159.csv)

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.9490305581239644	0.9817762237762242	1	0.9683333333333333	0.995	0.9985714285714291	0.9383333333333333

Overall Assessment

- **MCC (0.949)**: Near-perfect overall classification balance.
- **F1 Score (0.982)**: Excellent balance of precision and recall.
- **AUC (1)**: Perfect class separation.
- **TPR (0.968)** and **TNR (0.995)**: High sensitivity and specificity.
- **PPV (0.999)** and **NPV (0.938)**: Accurate predictions for both classes.

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_results100159.csv)



1. Genes with High Impact (MCC drops significantly):(**FCGR1A.1, BCL2, MYD88, P2RX7, OLFM4.1, HLA-DRA**)

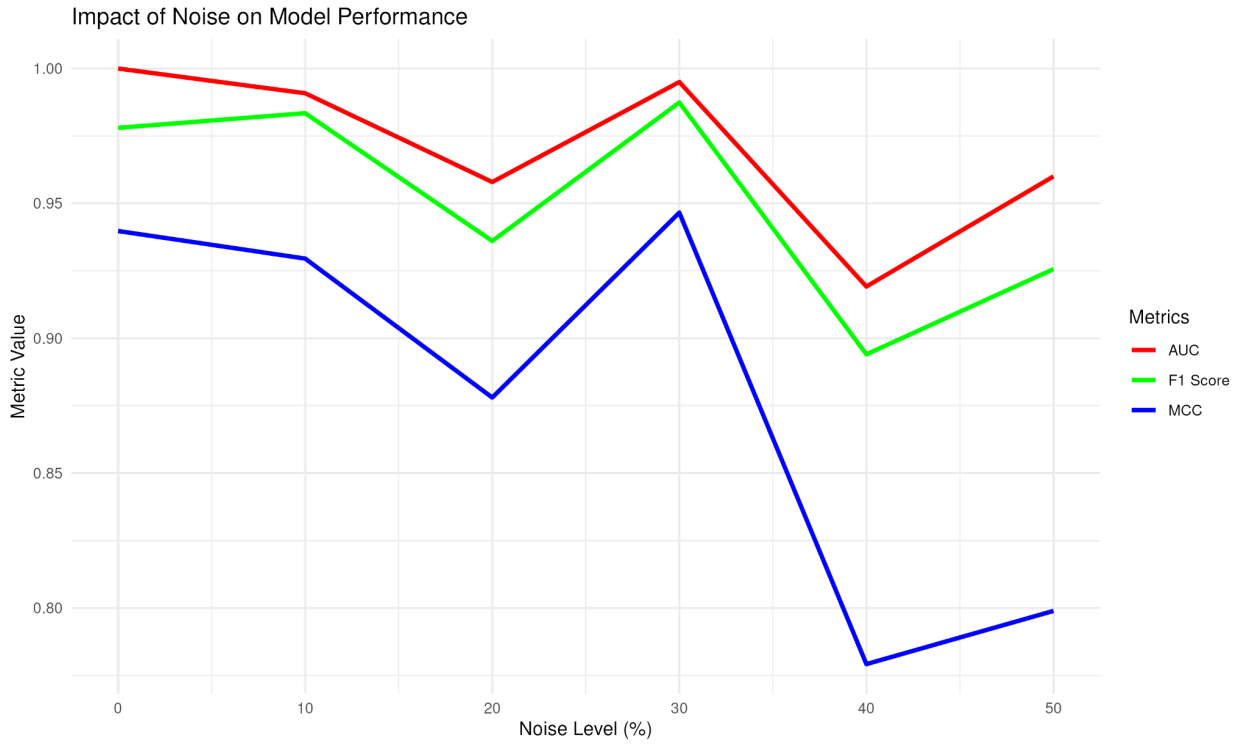
- Removing these genes causes a sharp decline in MCC, indicating they are critical for the model's classification performance.
- These are likely the most informative genes for predicting the outcome (e.g., sepsis).

2. Genes with Low Impact (Minimal change in MCC):(**HIF1A.1, PLAUR, NLRP3, IL1R2.1, TLR2, IL1B)**

- 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_results100159.csv)



Impact of Noise on Metrics:

- **AUC and F1 Score** are more robust to noise compared to MCC , 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** drop significantly, indicating that the model's ability to balance between positive and negative cases deteriorates under noisy conditions, especially for identifying negatives (controls).

The important problem in this dataset is the imbalancing of target levels. The train data includes 8 controls and 27 sepsis and test data includes 2 controls and 6 sepsis.
The resampling by SMOTE is done to make sure the results are reliable .

The ROSE resampling didn't work properly, maybe because of the small size.

Now here is the result for random forest after resampling by SMOTE:

Random Forest Model:

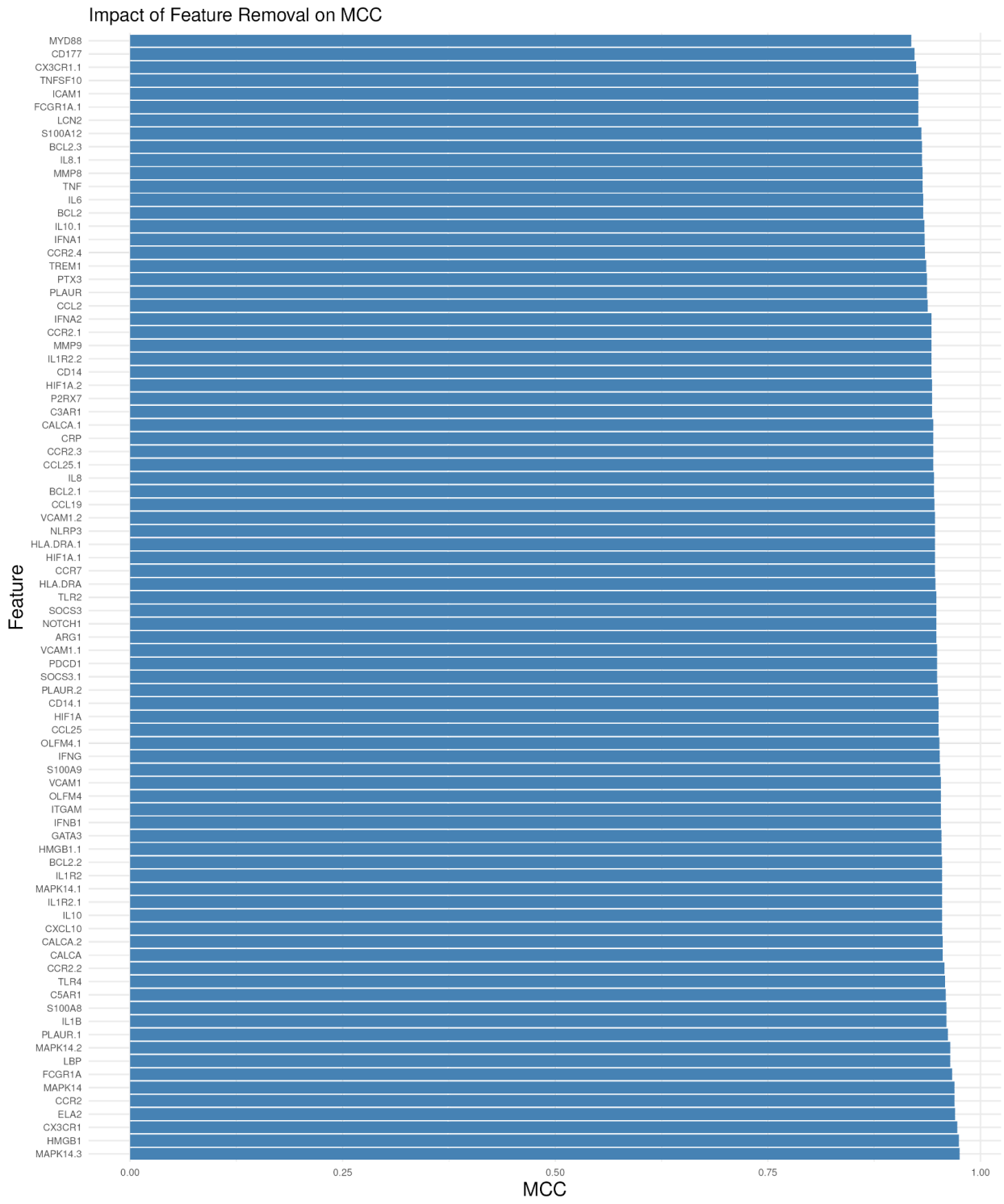
Random forest applied to the database 100 times by randomly splitting the dataset and resampling done by SMOTE. (repeated_splits_metrics_smote100159.csv)

Average Metrics
The average of 100 times repeated for each metric (MCC, F1 score, AUC, TPR, TNR,...) is calculated . (average_metrics_smote100159.csv)

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.927886003508775	0.973272727272727	1	0.951666666666667	1	1	0.906666666666667

There are slight changes in results, but overall results are almost the same.

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_smote100159.csv)



3. Genes with High Impact (MCC drops significantly):(**CD177, CX3CR1.1, MYD88, TNFSF10, ICAM1**)

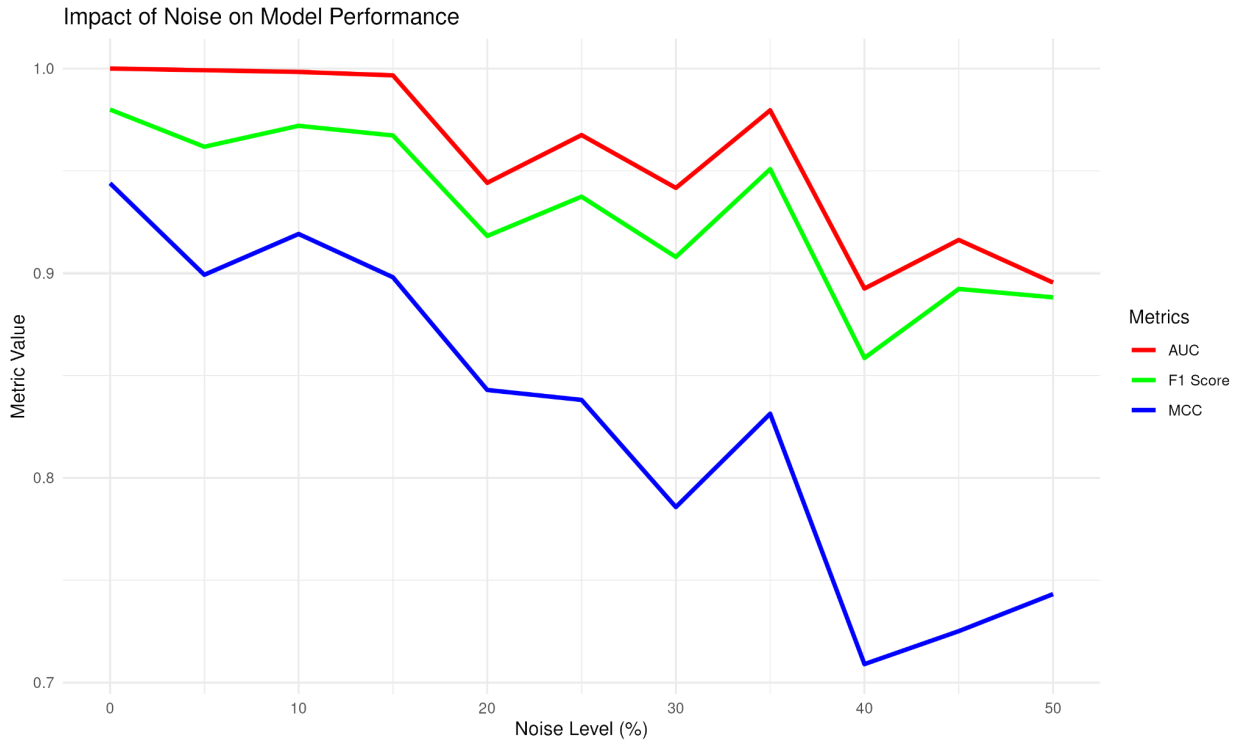
- Removing these genes causes a sharp decline in MCC, indicating they are critical for the model's classification performance.
- These are likely the most informative genes for predicting the outcome (e.g., sepsis).

4. Genes with Low Impact (Minimal change in MCC):(**MAPK14.3, HMGB1, CX3CR1, ELA2, CCR2, MAPK14)**

- 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_smote100159.csv)



Impact of Noise on Metrics:

- **AUC and F1 Score** are more robust to noise compared to MCC, 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** drops significantly, indicating that the model's ability to balance between positive and negative cases deteriorates under noisy conditions, especially for identifying negatives (controls).