Dataset:GSE32707
A microarray analysis involving whole blood samples isolated from critically ill patients in the medical intensive care unit at Brigham and Women's Hospital. Four groups of intubated subjects undergoing mechanical ventilation were recruited for the study: those with sepsis alone (Sepsis), those with sepsis + ARDS (se/ARDS), those with SIRS (SIRS), and those without sepsis, SIRS, or ARDS (untreated). Blood was obtained from patients on the day of admission (day 0) and 7 days later. RNA was isolated from the whole blood samples and microarrays were prepared to determine differential gene expression between the four groups.
Platform: GPL10558 (Illumina HumanHT-12 V4.0 expression beadchip).

Study Design

- **Focus**: The dataset involves critically ill patients and examines inflammasome-regulated cytokines, which are crucial for lung injury.
- **Groups**: There are four groups of patients:
    - Patients with sepsis alone (*Sepsis*).
    - Patients with sepsis and acute respiratory distress syndrome (ARDS) (*Sepsis + ARDS*).
    - Patients with systemic inflammatory response syndrome (SIRS) but no sepsis.
    - Patients without sepsis, ARDS, or SIRS (*Untreated*).
- **Samples**: RNA was obtained from the whole blood of these patients on:
    - **Day 0**: The day of admission.
    - **Day 7**: Seven days later.

Dataset was already normalized. Three different datasets are extracted from the data.(GSE32707code.R)
1. filtered_sepsD0_GSE32707.csv -> sepsis patients for day 0 of admission. 30 sepsis patients and 34 controls. (64 rows and 94 columns)

2. filtered_sepsD7_GSE32707.csv -> sepsis patients for day 7 of admission. 28 sepsis patients and 34 controls.(62 rows and 94 columns)

3. filtered_sepsis_all_GSE32707.csv -> sepsis patients for day 0 and 7 of admission. 58 sepsis patients and 34 controls.(92 rows and 94 columns)

In this dataset the control group is not healthy control , they are Patients without sepsis, ARDS, or SIRS.

Presented-genes: 55
Missing-genes : 0

Some filtered genes of interest in the dataset have duplication, there are more than one column for one gene. Based on (Towards a potential pan-cancer prognostic signature for gene expression based on probesets and ensemble machine learning)(https://doi.org/10.1186/s13040-022-00312-y) all the column related to each probe id are considered.

Sepsis patients of day0:

Random Forest Model

Here is the result for patients from day 0 . Random forest applied to the database 100 times by randomly splitting the dataset.(repeated_splits_metrics.csv)

Average Metrics

The average of 100 times repeat for each metric is calculated.(average_metrics.csv)

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.46440254 8356378	0.70489346 7643468	0.81819444 4444444	0.69166666 6666667	0.75166666 6666667	0.76161111 1111111	0.72643253 968254

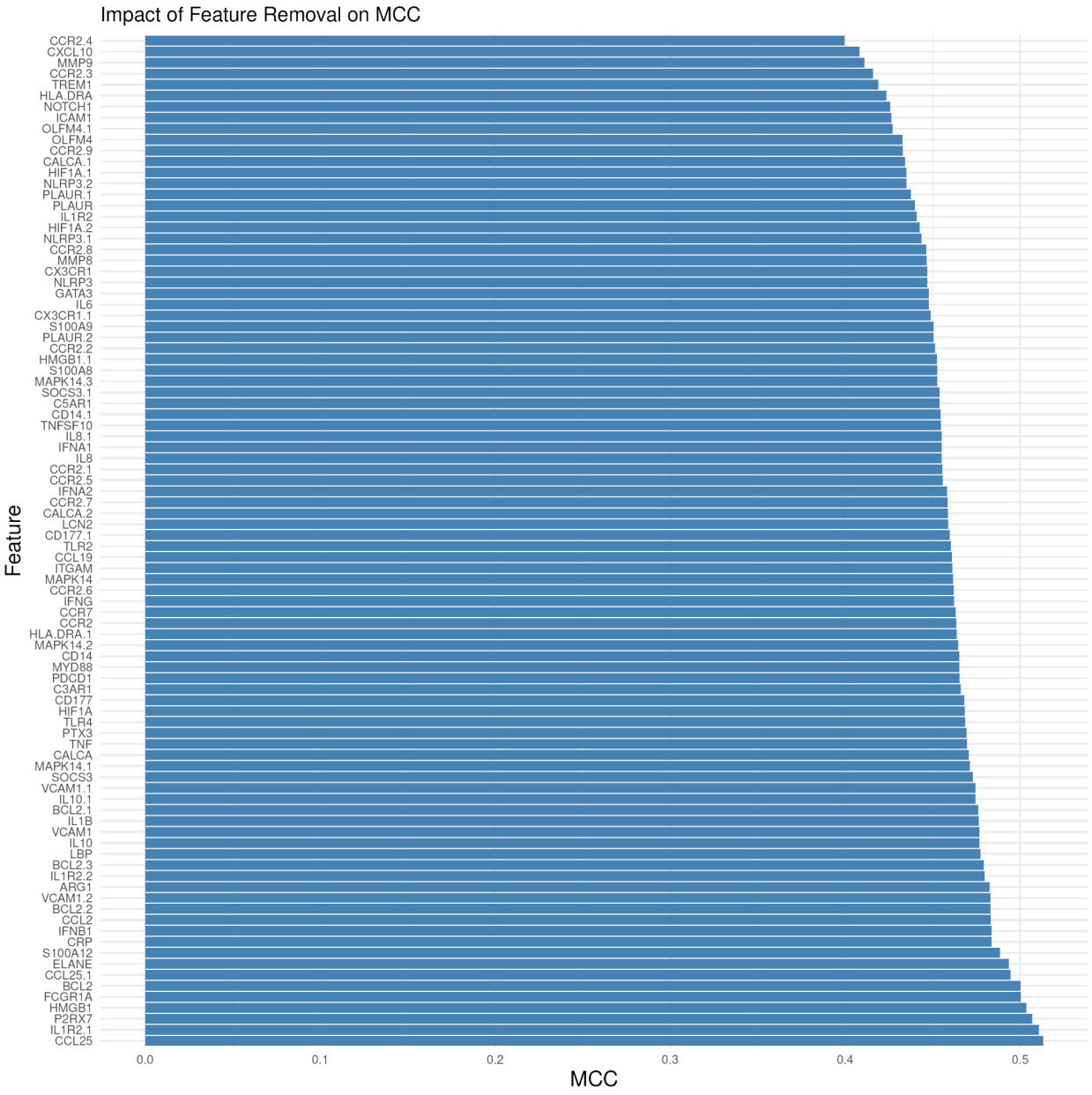
The model performs moderately well, with **good discrimination ability (AUC = 0.8182)** and a fair balance between precision and recall (F1 = 0.7049).

However, the **MCC (0.4644)** suggests there is room for improvement in the model's overall predictive quality.

Sensitivity (TPR) is slightly lower than **Specificity (TNR)**, indicating the model is slightly better at identifying negative cases than positive cases.

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.csv)



1. Genes with High Impact (MCC drops significantly):(**OCR2.4,CCR2.3, CXCL10, MMP9, NOTCH1, HLA.DRA, TREM1**)

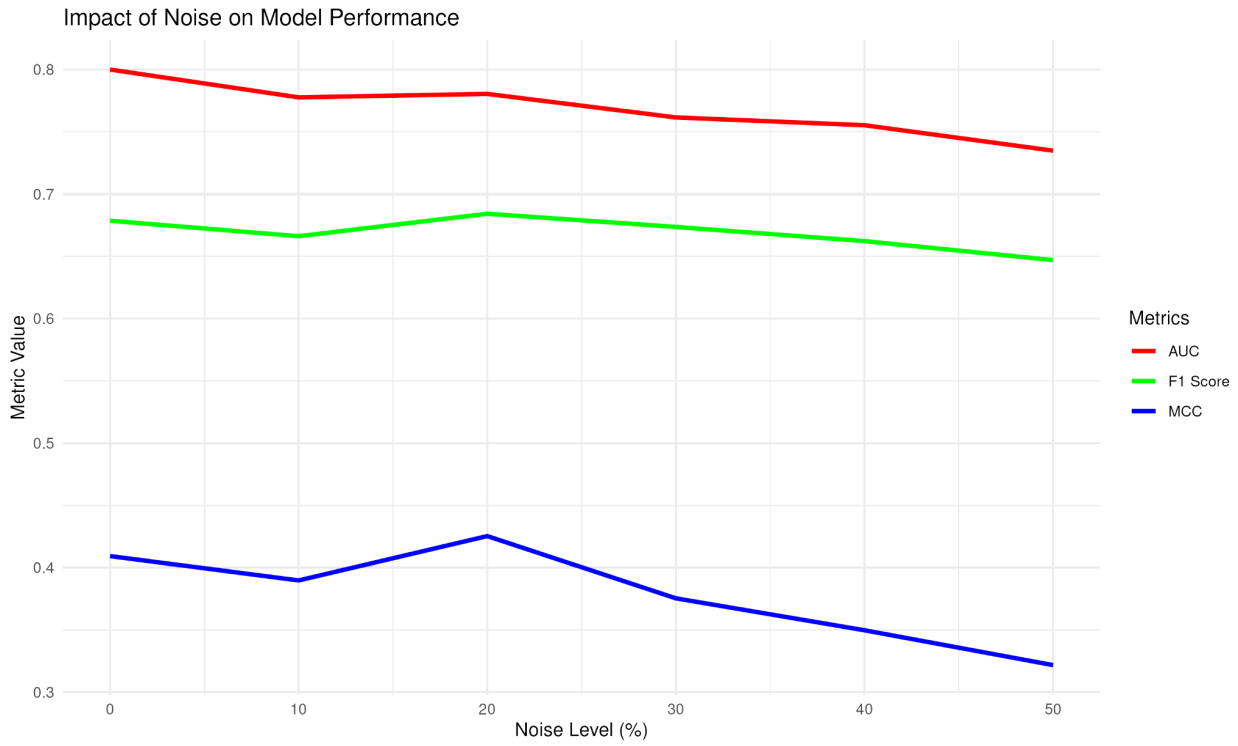
- Removing these genes causes a sharp decline in MCC, indicating they are critical for the model's classification performance.
- These are likely the most informative genes for predicting the outcome (e.g., sepsis).

2. Genes with Low Impact (Minimal change in MCC):(**CCL25, IL1R2.1, P2RX7, HMGB1, FCGR1A, BCL2)**

- 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(impact_of_noise_on_model_performance.png) and table(sanity_check_results.csv) represent how the performance of your model changes as increasing levels of noise are added to the dataset.



At Noise Level 0%:

- The model performs well, with MCC, F1, and AUC showing strong classification capabilities. The metrics indicate good sensitivity (TPR), specificity (TNR), and balanced predictions.

As Noise Increases:

- The metrics steadily worsen, with noticeable declines in MCC, F1, and AUC. This reflects the model's reduced ability to accurately classify instances as the data becomes increasingly corrupted.
- Precision (PPV) and recall (TPR) both decline, showing the model struggles with positive cases. Similarly, TNR and NPV decrease, indicating issues with negative cases.

Sepsis patients of day7:

Random Forest Model

Here is the result for patients from day 7 . Random forest applied to the database 100 times by randomly splitting the dataset.(repeated_splits_metrics.csv)

Average Metrics

The average of 100 times repeat for each metric is calculated.(average_metrics.csv)

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.52653096 2388602	0.73391347 5413475	0.83116666 6666667	0.74	0.77	0.75842063 4920635	0.78665476 1904762

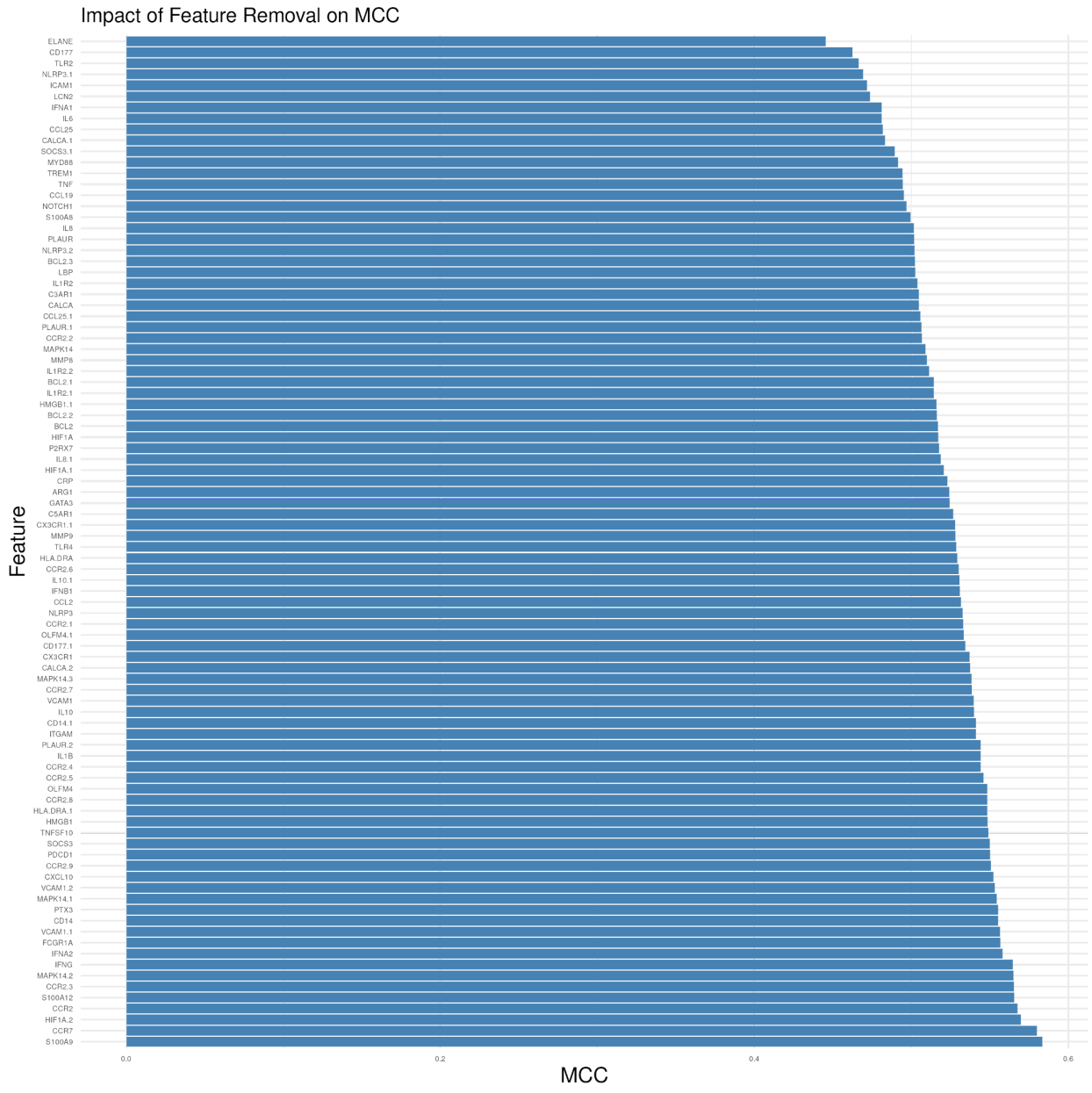
The model performs moderately well, with **good discrimination ability (AUC = 0.8311)** and a fair balance between precision and recall (F1 = 0.7339).

However, the **MCC (0.4644)** suggests there is room for improvement in the model's overall predictive quality.

Sensitivity (TPR) is slightly lower than **Specificity (TNR)**, indicating the model is slightly better at identifying negative cases than positive cases.

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.csv)



Genes with High Impact (MCC drops significantly):(**ELANE, CD177, TLR2, NLRP3.1, ICAM1, LCN2, IFNA1**)

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

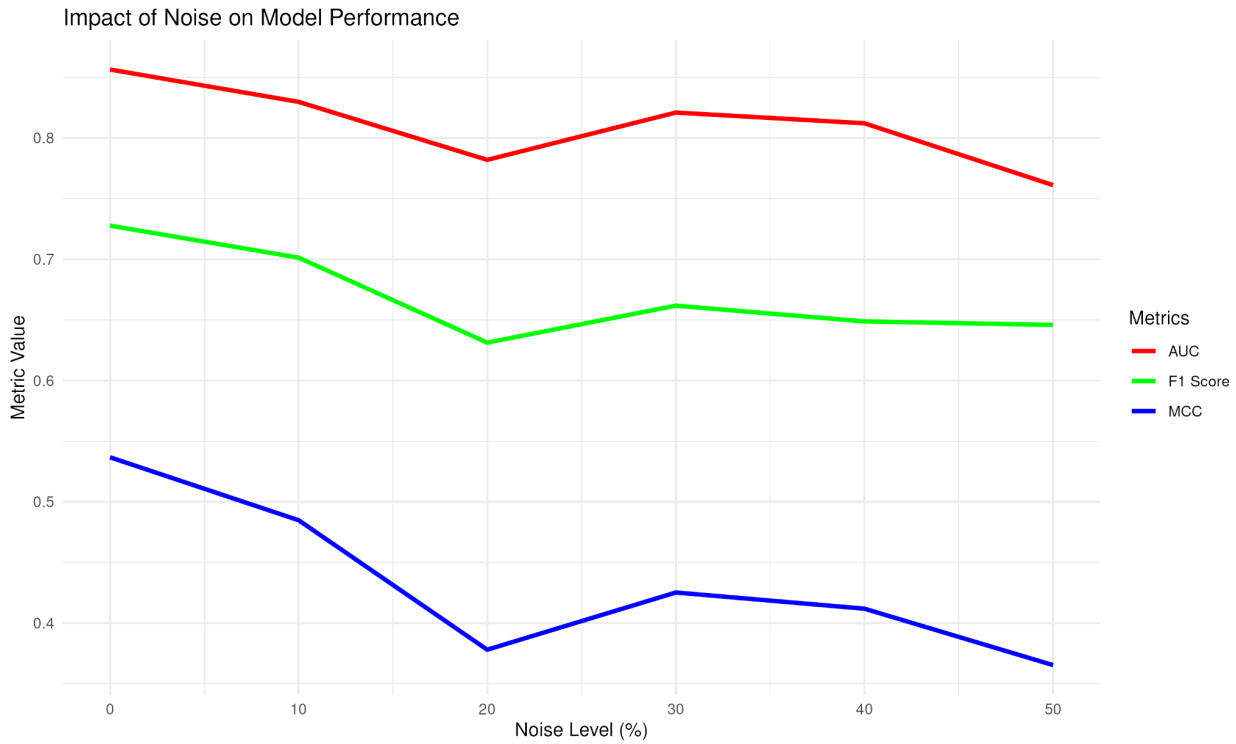
- These are likely the most informative genes for predicting the outcome (e.g., sepsis).

Genes with Low Impact (Minimal change in MCC):(**S100A9, S100A12, CCR2, CCR7, HIF1A.2, CCR2.3)**

- 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(impact_of_noise_on_model_performance.png) and table(sanity_check_results.csv) represent how the performance of your model changes as increasing levels of noise are added to the dataset.



At Noise Level 0%:

- This model also performs well, with MCC, F1, and AUC showing strong classification capabilities. The metrics indicate good sensitivity (TPR), specificity (TNR), and balanced predictions.

As Noise Increases:

- The metrics steadily worsen, with noticeable declines in MCC, F1, and AUC. This reflects the model's reduced ability to accurately classify instances as the data becomes increasingly corrupted.
- Precision (PPV) and recall (TPR) both decline, showing the model struggles with positive cases. Similarly, TNR and NPV decrease, indicating issues with negative cases.