Dataset: **GSE211210**

This dataset, identified as GSE211210, explores the molecular mechanisms underlying sepsis pathogenesis, focusing on the role of human ALKBH8. Expression profiling by high-throughput sequencing (RNA-seq).

Platform: GPL24676 (Illumina NovaSeq 6000).

Study Design:

- The study includes **10 patients** with acute sepsis and **10 healthy controls**.
- RNA was extracted from blood samples of sepsis patients (**n = 5**) and healthy controls (**n = 5**) for RNA-seq analysis.

The dataset had already fpkm-normalized data, then we prepared it for downstream analysis. (sepsis_dataGSE211210.csv). (rows = 10 and column = 57)

Present genes : 55
Missing genes : 0

Random forest model:

Then Random forest applied to the dataset that target label considered as factor.(Sepsis and Control)

100 randomly split was done for random forest (repeated_splits_metrics21120.csv).
Average metrics calculated for all MCC, F1, AUC,... and saved in (average_metrics211210.csv) file. Here is the results:

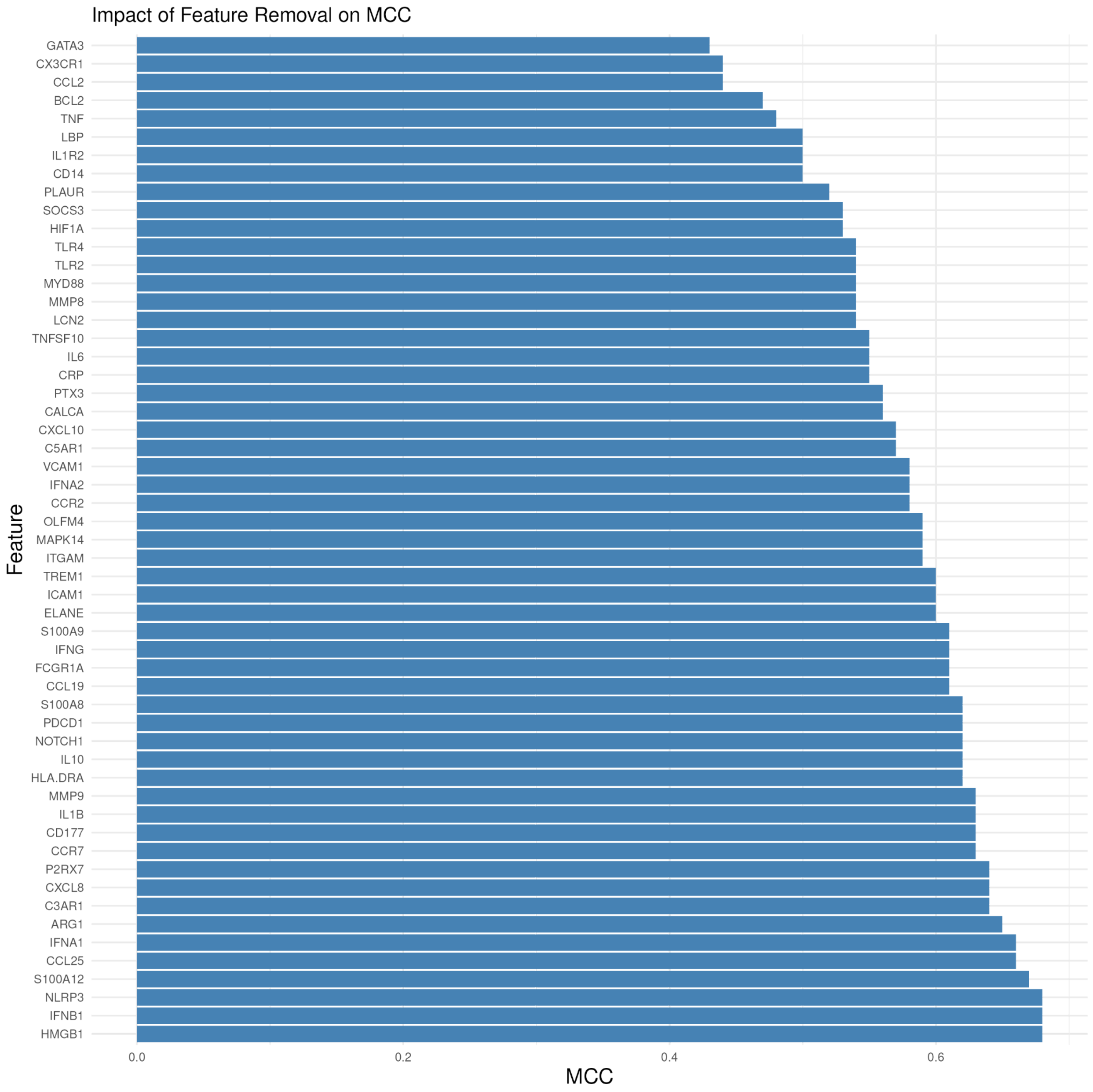
Average metrics:

MCC	F1	AUC	TPR	TNR	PPV	NPV
0.57	0.585858585858586	0.59	0.58	0.58	0.58	0.58

Feature removal analysis:

To investigate the importance of each gene in the prediction of the model , in each 100 iterations one feature was removed from the dataset to evaluate the mentioned metrics.(feature_removal_results211210.csv)

The plot that is created by results of feature removal:(feature_removal_mcc_plot211210.png)



Most important genes based on the plot that we see are the genes that by removing from the dataset cause more drop in MCC value. (**GATA3, CX3CR1, CCL2, BCL2**).

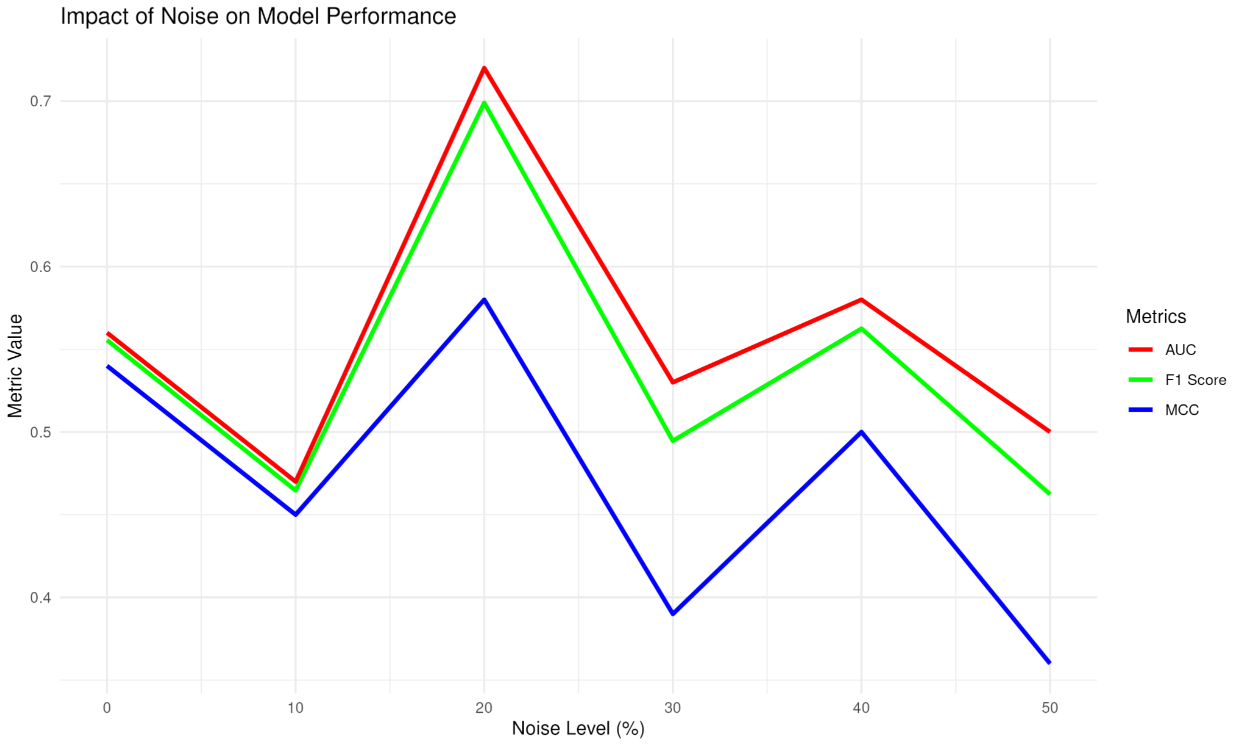
Less important genes in the prediction performance of the model are those that their removal does not affect the MCC value significantly. (**HMGB1, IFNB1, NLRP3, S100A12 and CCL25**)

Sanity check:

Sanity check has been done on the model to evaluate its performance . By adding more noise step by step to the dataset in each iteration to confirm if the MCC, F1 and AUC value will decline or not.

The noise was added to the dataset in 5 different steps as 10%, 20%, 30%, 40% and 50%.(sanity_check_results211210.csv).

Here is the plot that shows how the calculated metrics will change by increasing noise on the dataset. (impact_of_noise_on_model_performance211210.png).



The problem was many 0 values in some features of the dataset.

So we normalized the raw data of the dataset to make sure it's properly normalized.

The Deseq2 was used for normalization and then annotation data was used to map genes symbols and filter them and prepare them for downstream analysis. ((sepsis_Deseq_normGSE211210.csv)). But it still had 4 or 5 features with 0 values.

The results were the same but there was an issue in results, because the dataset is too small. In the train data we had just 4 sepsis and 4 controls and in test data we had only 1 for each of them which makes the result unreliable.

The next step was try to make synthetic remapping methods like ROSE or SMOTE but it didnt work properly because of small dataset and we got these errors each time

```
[1] "SMOTE balancing failed. Consider upSample or an alternative method."
[1] "ROSE balancing failed. Using upSample as fallback."
```

Then tried to use an upsample, but results were still the same as before in the number of samples for each label in train and test data.

Here is the result after using Upsample for 100 random split:

```
confusion

Confusion Matrix and Statistics


          Reference

Prediction Control Sepsis

   Control       1      0

   Sepsis        0      1


               Accuracy : 1

                 95% CI : (0.1581, 1)
```

```
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.25

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.5
         Detection Rate : 0.5
   Detection Prevalence : 0.5
      Balanced Accuracy : 1.0

       'Positive' Class : Sepsis

> table(train_data_balanced$Label)

Control  Sepsis
      4       4
> table(train_data$Label)

Control  Sepsis
      4       4
> table(test_data$Label)
```

```
Control  Sepsis

      1       1
```