Dataset: GSE185263

The dataset GSE185263 comprises gene expression profiles derived from whole blood RNA sequencing of **348 patients** across four emergency rooms and one intensive care unit. This dataset was designed to explore transcriptional changes in sepsis patients, assess disease severity, predict organ dysfunction, mortality, and uncover specific endotypes and mechanisms related to sepsis.

Data Preparation and Processing:

Two labeled datasets have been generated: (rows 392, cols 57)

Sepsis vs. Healthy Dataset: Includes samples labeled as either Sepsis or Healthy Controls for sepsis diagnosis.(sepsis_labeled_dataGSE185263.csv).

Sepsis = 348 —- healthy = 44.

Prognostic Dataset: Includes samples labeled as Survived or Deceased for assessing the prognosis of sepsis patients.(sepsis_prognos_labeledGSE185263.csv)

Survived = 293 —- died = 52 —- NA = 47.

All the prepared files are in the Dataset directory with corresponding GSE number.

Genes of interest have been identified and filtered based on their Ensembl Gene IDs and Gene Symbols. All genes of interest were present in the dataset and successfully searched using both Ensembl Gene IDs and Gene Symbols through BioMart.

Presented-genes: 55

Missing-genes :0

Normalization of raw count data

The dataset consists of raw read counts for each gene across samples. These counts represent the number of sequencing reads mapped to each gene, which are not directly comparable between samples without normalization.

Variance Stabilization Normalization (VSN):The VSN method was applied to normalize the raw counts. VSN uses a variance stabilization transformation that makes the variance roughly constant

across the range of mean intensities. This approach helps to reduce heteroscedasticity in the data, which can improve downstream statistical analyses.

The process involves:

Converting the raw counts into a numerical matrix, excluding metadata (e.g., Ensembl Gene IDs).

Applying the justvsn function from the vsn R package to normalize the data.

Verifying the normalization fit using the meanSdPlot function to ensure that the relationship between mean and standard deviation of gene expression is stabilized.

Random forest Model:

The random forest test is done on (sepsis_labeled_dataGSE185263.csv) dataset. and the target label was considered as factor. (RFlabel-GSE185263code.R)

here is the results:

average metrics results:

After 100 repeated random forest by randomly splitting the data(repeated_splits_metrics.csv), the average metrics was :

"MCC"	F1	AUC	TPR	TNR	PPV	NPV
0.792568919 96504	0.979106278 871455	0.988940217 391304	0.984492753 623188	0.77125	0.973971493 263954	0.863718253 968254

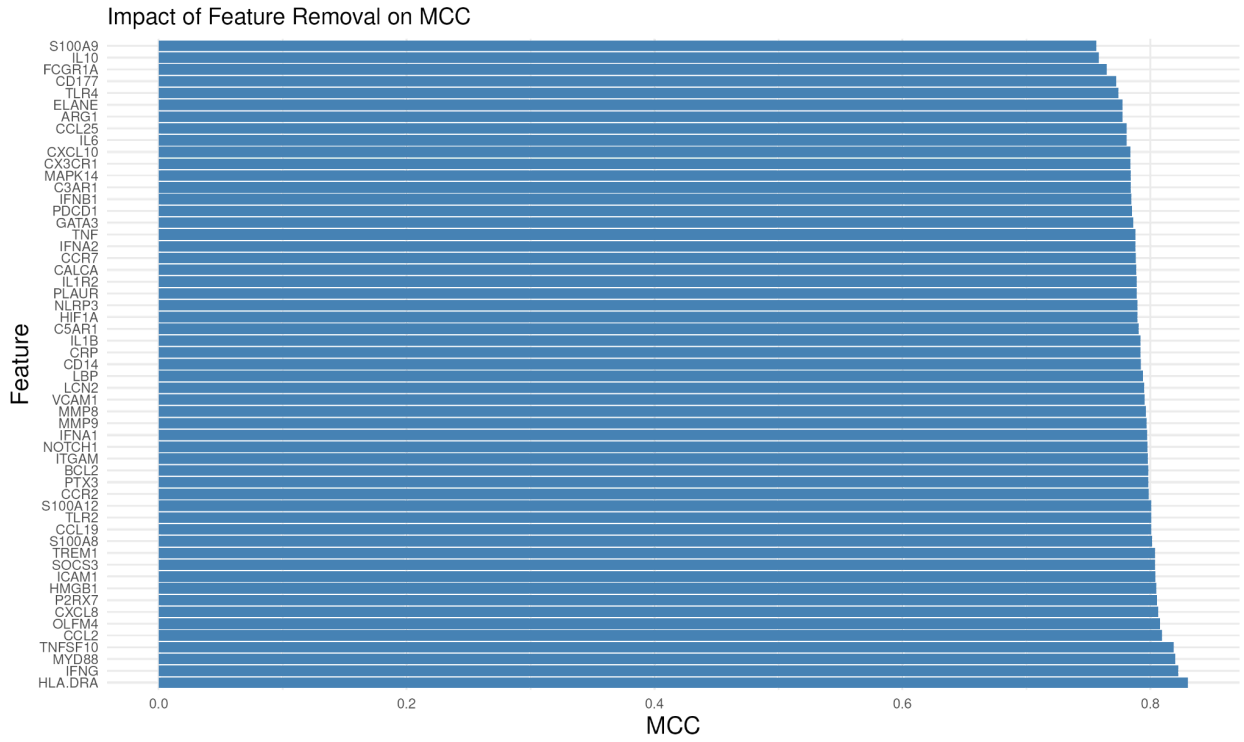
Feature removal random forest:

Shows by removing which column that is represnts one gene of interest the performance for prediction will drop significantly.(feature_removal_results.csv)

feature removing plot:

X-Axis: The values along the x-axis represent the performance difference (using MCC) when a specific feature is removed from the model.

Y-Axis: The features (genes) are listed. Features at the top show a greater drop in performance upon removal, while features at the bottom show a smaller drop in performance.



Top Features (S100A9, IL10, FCGR1A, CD177, TL4):

These are likely key biomarkers and should be prioritized in biological interpretation and downstream analysis. They are central to the model's predictions and could be explored further for biological relevance.

Middle Features (CCR7, PTX3, CALCA, ITGAM):

These features are moderately important and could complement the top features. They might also contribute indirectly through feature interactions.

Bottom Features (MYD88, HLA.DRA, IFNG, TNFSF10, CCL2):

These features might not be crucial for prediction in this context. Consider removing them to simplify the model or explore their redundancy with other features.

Sanity check

Noise is artificially introduced into the dataset, and the model's performance metrics are observed as the noise level increases.(sanity_check_results.csv)

Noise Level 0: Represents the original data with no added noise (baseline performance).

Noise Levels 10–50: Increasing levels of noise are added to the input data, simulating scenarios with reduced data quality or signal interference.

Metrics (MCC, F1, AUC, TPR, TNR, PPV, NPV):

MCC (Matthews Correlation Coefficient): Measures the quality of predictions (balanced for imbalanced datasets).

F1 Score: Balances precision and recall.

AUC (Area Under the Curve): Measures the ability to distinguish between classes.

TPR (True Positive Rate): Sensitivity (recall for positives).

TNR (True Negative Rate): Specificity (recall for negatives).

PPV (Precision for positives): Probability that a positive prediction is correct.

NPV (Precision for negatives): Probability that a negative prediction is correct.

"Noise_Level","MCC","F1","AUC","TPR","TNR","PPV","NPV"

0,0.787146199317093,0.978701046312937,0.989827898550725,0.985072463768116,0.7587
5,0.972690193393951,0.869672438672439

10,0.774733683746892,0.977824521118772,0.985471014492754,0.986231884057971,0.732
5,0.969842927893636,0.877285714285714

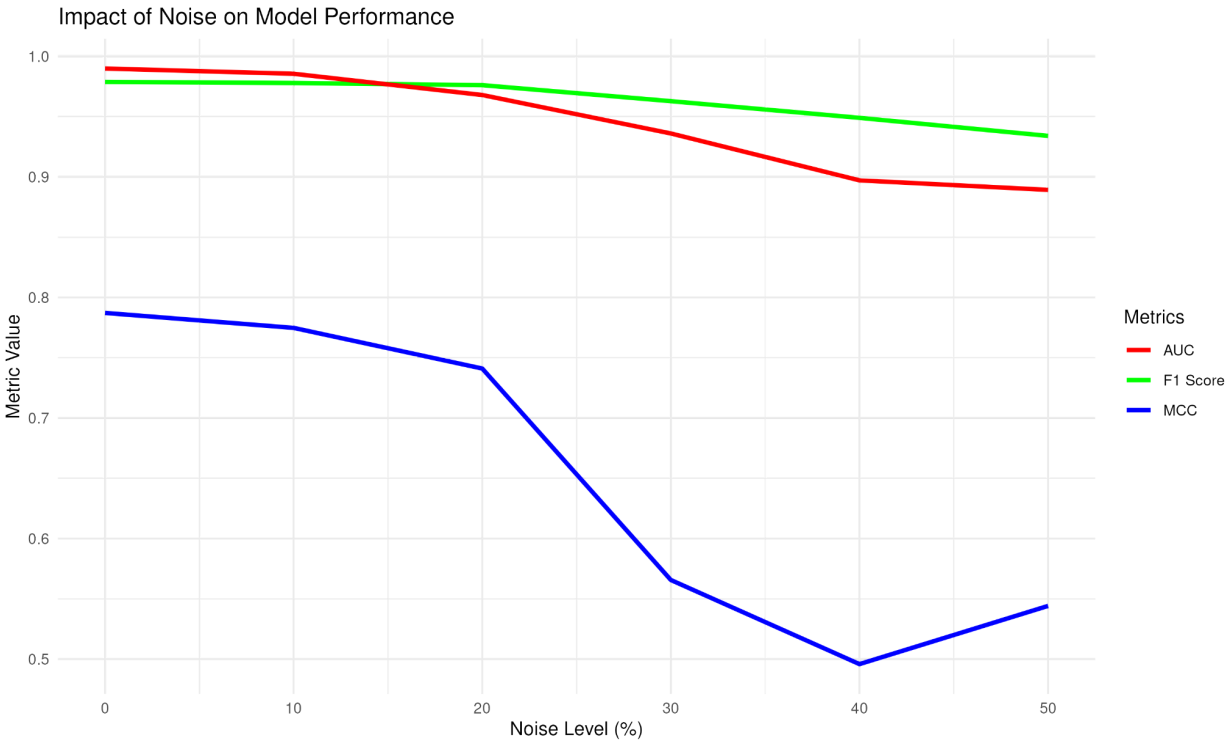
20,0.741025278620514,0.976029239802289,0.96786231884058,0.991884057971015,0.6487
5,0.960935688317965,0.912981240981241

30,0.565663317123334,0.962752982763122,0.935969202898551,0.987536231884058,0.447
5,0.939441067688016,0.822511904761905

40,0.495896101953109,0.94892690181517,0.897083333333333,0.978985507246377,0.3537
5,0.920831466864912,0.81527380952381

50,0.544160688667241,0.933970303115782,0.889184782608696,0.962753623188406,0.391
25,0.907020314035807,0.857488095238095

Impact of Noise:



Moderate Noise (10–20): The model shows minor declines in MCC, F1, and AUC.

High Noise (30–50): Significant degradation in performance, particularly in MCC, AUC, and TNR. This suggests the model struggles with highly noisy conditions.