

ABSTRACT

Infectious diseases, notably sepsis, impose a global health burden and require careful analysis of transcriptomics data to predict mortality outcomes. Current studies lack optimal machine learning models that take into account demographics, infection, and severity of the disease to analyze transcriptomics data to identify and estimate mortality rates. Here we have assembled and analyzed microarray datasets by 5 Machine Learning (ML) algorithms to estimate mortality rates of sepsis. We have also benchmarked these performance of ML models using various metrics to choose the best predictive model, and found that Light GBM model (Light Gradient Boosting) predicted mortality at the highest accuracy of 75%.

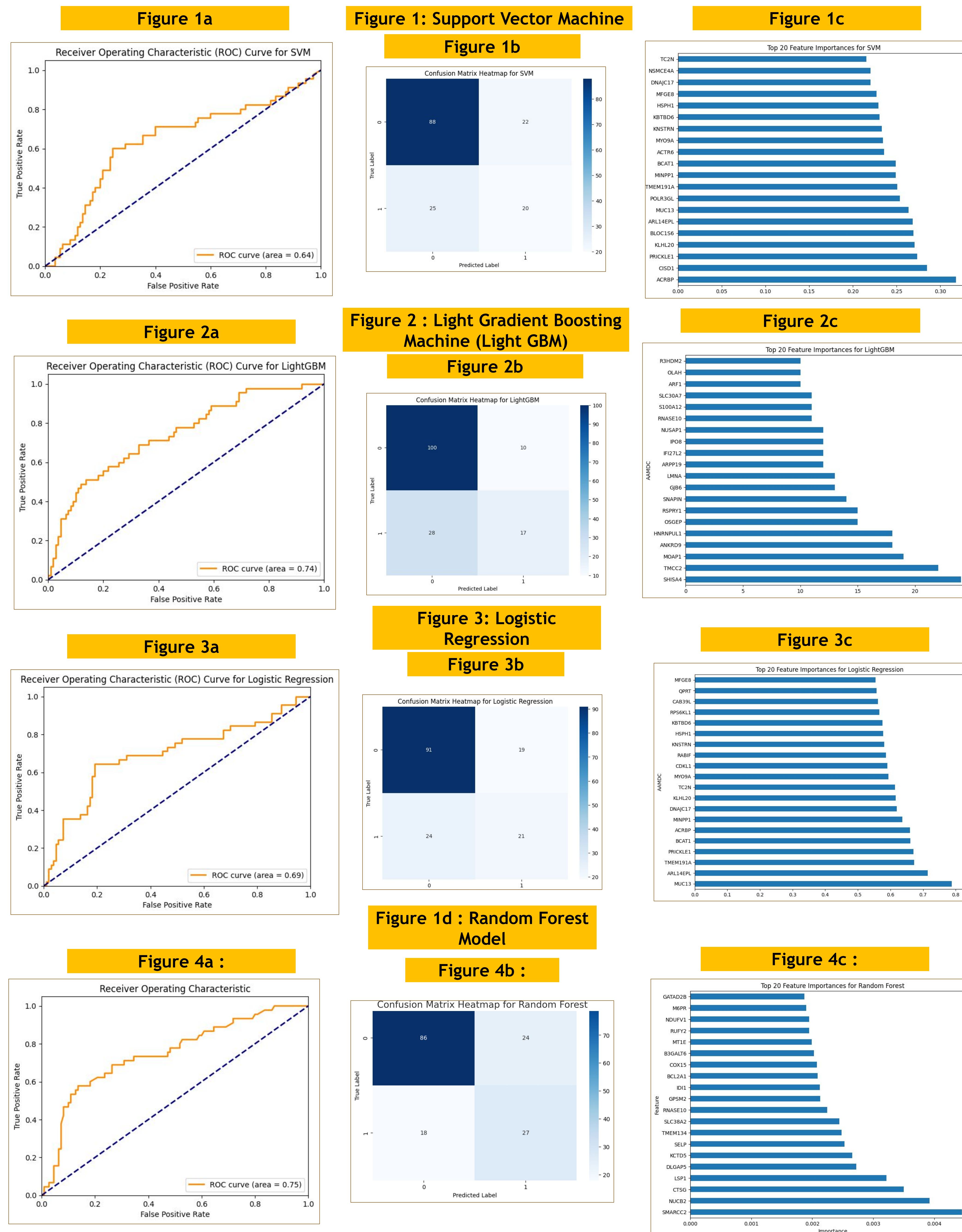
INTRODUCTION

- Sepsis is one of the rising concerns for global health deterioration that still lacks comprehensive transcriptomics data analysis to aid in the prediction of mortality outcomes. [1]
- The ability of machine learning models to utilize complex datasets have enabled pattern recognition and classification, broadening application in bioinformatics and medicine. [2]
- In this study, we analyzed microarray expression datasets taken from public repositories utilizing seven microarray gene expression datasets which encompassed more than 3000 unique genes, to create a feature set for 772 patients, with 540 samples representing survival outcomes (positive outcome) and 232 samples representing mortality (negative outcome) and conducted predictive analysis on Machine learning models.

METHODS

- Microarray expression datasets were taken from public repositories namely NCBI (National Center for Biotechnology Information) GEO (Gene expression Omnibus) [3] and EBI (European bioinformatics Institute) [4]
- Processed datasets included approximately 7000 genes, from which 3153 common genes across all 7 experiments were selected.
- Survival outcomes of these experiments were also extracted from the metadata manually for these experiments to form and were employed into ML models to accelerate predictions.
- To address the class imbalance in the dataset, we applied the Synthetic Minority Over-sampling Technique (SMOTE).
- We used 80% of the data for training and 20% for testing. We employed four Machine Learning models—Support Vector Machine (SVM), Light Gradient Boosting Machine (Light GBM), Logistic Regression, and Random Forest—to evaluate mortality outcomes and conduct benchmarking studies.

RESULT



RESULTS AND DISCUSSION

- Our analysis indicates that, among the four models, Light GBM achieved the highest accuracy of 75% for estimating mortality outcomes of sepsis, with an AUC of 0.74 (Figure 2a). This was closely followed by Random Forest (Figure 4a) and Logistic Regression (Figure 3a), each with an accuracy of 72% and an AUC of 0.69. SVM exhibited the lowest accuracy of 70% but had an AUC of 0.74.
- Genes that show to have the most impact on survival outcome from Light GBM model was SHISA4 (Figure 1c) which codes for transmembrane-adaptor protein. [5]
- Some of the common genes of importances predicted by SVM and Logistic regression models were TC2N, MFGE8 and HSPH1 amongst others. These genes and changes in their expression levels were found to be associated with different severities of sepsis [5]

REFERENCES

- Bauer M, Gerlach H, Vogelmann T, Preissner F, Stiefel J, Adam D. Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019: results from a systematic review and meta-analysis. Crit Care. 2020 May 19;24(1):239.
- Kasula, B. Y. (2019). Enhancing Classification Precision: Exploring the Power of Support-Vector Networks in Machine Learning. International Scientific Journal for Research, 1(1).
- National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] available from: <https://www.ncbi.nlm.nih.gov/geo/>.
- Supporting data are available in the BioStudies database (<http://www.ebi.ac.uk/biostudies>) under accession number : E-MTAB-4421.
- Chen, L., Chen, K., Hong, Y. et al. The landscape of isoform switches in sepsis: a multicenter cohort study. Sci Rep 12, 10276 (2022)

CONTACT



Web-page: <https://mangul-lab-usc.github.io/>

Sanskriti Sharma

Department of Pharmaceutical Sciences, School of Pharmacy, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033, USA
Email: susharma@usc.edu

Nitesh Kumar Sharma

Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033, USA
Email: prince26121991@gmail.com (NKS)

Serghei Mangul

Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033, USA
Email: serghei.mangul@gmail.com