Leveraging Admission and Serial Transcriptomic Data to Improve Delirium Prediction in COVID-19 Patients

Shreya Vora, Grace Xie, Garima Upadhyay, Matthew Ho, Audrey Su

with Dr. Neel Singhal (UC San Francisco, Dept of Neurology) & UC Berkeley CDSS Discovery Program

BACKGROUND

Delirium, characterized by sudden changes in awareness, attention, and cognition, affects 18-35% of hospitalized patients and is linked to increased mortality and longer hospital stays. Trials like AID-ICU and Mind-USA found no significant benefits of haloperidol or ziprasidone over placebo, underscoring the need for better tools to predict delirium risk.

Predictive models that integrate demographic and genetic features could allow for earlier intervention. Our research utilizes one of the most comprehensive datasets of serial transcriptomics in COVID-19 patients to explore delirium risk. By analyzing gene expression trajectories with admission data, we aim to find early predictors of delirium, advancing preventative approaches to patient care.

OBJECTIVE & DATA

We aimed to evaluate the utility of admission demographic and transcriptomic data, alongside serial gene expression trajectories, in predicting delirium risk among patients, specifying the demographic and genetic factors that were most associated with the development of delirium.

We utilized a dataset comprising transcriptomic data from 258 patients hospitalized with COVID-19 over multiple days. The dataset includes gene expression measurements for 13,107 genes, providing a comprehensive view of serial transcriptomic changes during hospitalization. In addition to the gene expression data, several demographic and clinical variables were collected. For the purposes of our analysis, we focused on demographic features available at the time of admission, specifically Age at Admission, Sex at Birth, Race, Hispanic Ethnicity, SOFA score, and WHO Scale. This approach allowed us to investigate early predictive indicators while leveraging the extensive gene expression data to explore potential molecular correlates of clinical outcomes.

CONCLUSION

Demographic data, admission transcriptomic data, and serial transcriptomic data were used to develop a series of COVID related delirium predictor models that aim to help with delirium diagnosis. The models created were able to reach increasingly impressive accuracies with the incorporation of diverse data collected from patient cohorts. Admission transcriptomics and serial transcriptomics combined models tended to perform better than admission transcriptomics models alone, which tended to perform better than pure demographic data based models. From a clinical perspective, model selected genes provide insight to the biological nature of COVID related delirium and contribute towards discovering a cohort of predictive blood-based biomarkers that will aid in early delirium diagnosis and treatment.

FUTURE WORK

While our research offers valuable insights, several limitations must be addressed. With only 25 of 258 patients experiencing delirium, our findings may lack generalizability and be biased toward predicting "no delirium." Although one of the largest datasets of its kind, the small sample size made model results sensitive to train-test splits. Inconsistent serial gene expression data further reduced usable samples, limiting the robustness and broader clinical applicability of our findings.

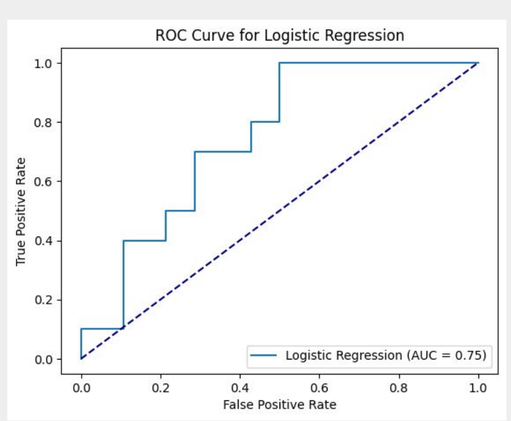
Future work could address these limitations by using larger, more balanced datasets and integrating additional data sources, such as medical imaging, clinical laboratory results, or EEG data, to improve model accuracy. Developing a risk scoring system based on identified predictors could make results more actionable, while exploring underlying biological mechanisms of delirium could reveal new therapeutic targets.

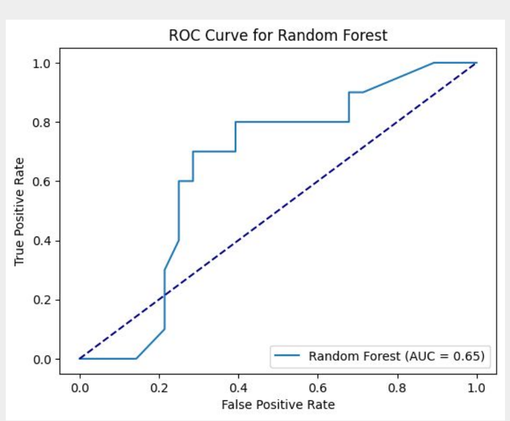
METHODS & ANALYSIS

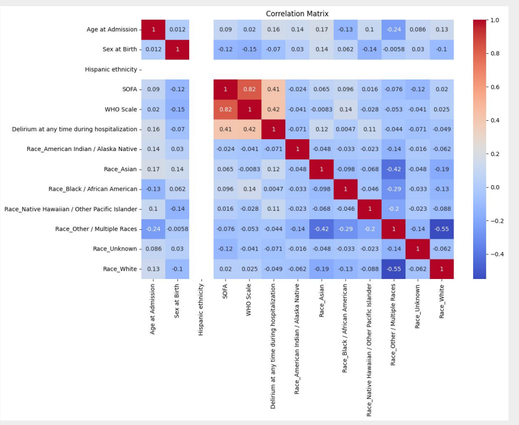
To start, we preprocessed the data by imputing missing values in numeric columns with the mean and one-hot encoding categorical variables to ensure a clean dataset for predictive analysis. Key demographic features available at admission–Age at Admission, Sex at Birth, Race, Hispanic Ethnicity, SOFA score, and WHO Scale–were used as inputs for various machine learning models, including Logistic Regression, Random Forest, LightGBM, XGBoost, PCA-transformed Logistic Regression, and K-Nearest Neighbors (KNN).

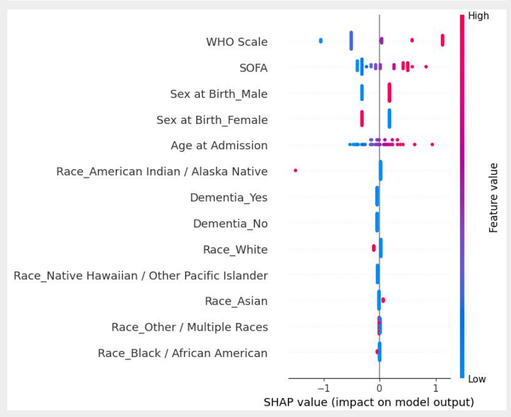
DEMOGRAPHIC

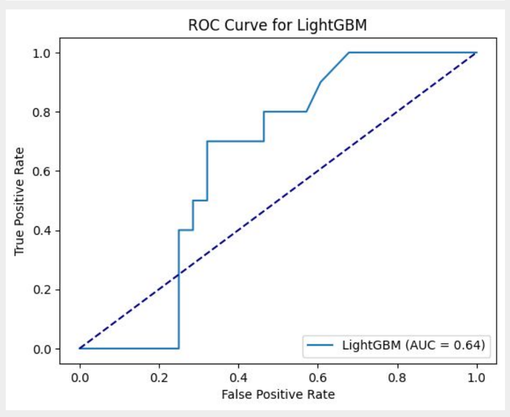
PCA-transformed Logistic Regression achieved the highest ROC accuracy of 0.75 and model accuracy of 0.68, while the simpler Logistic Regression model being marginally less accurate. KNN with selected features (Age at Admission, SOFA score, WHO Scale) demonstrated improved model accuracy of 0.73, compared to the model accuracy of 0.65 for KNN without narrow feature selection. SHAP plots and a correlation matrix revealed SOFA score, WHO Scale, and Age at Admission as the most influential predictors. However, demographic-only models showed limited predictive power, suggesting the need to integrate transcriptomic data for improved accuracy. This highlights the potential of a multimodal approach to better identify patients at risk for delirium.

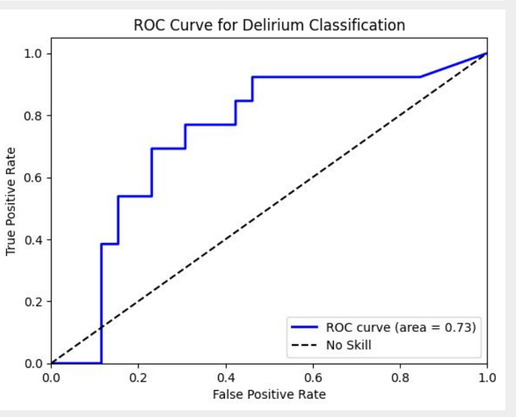


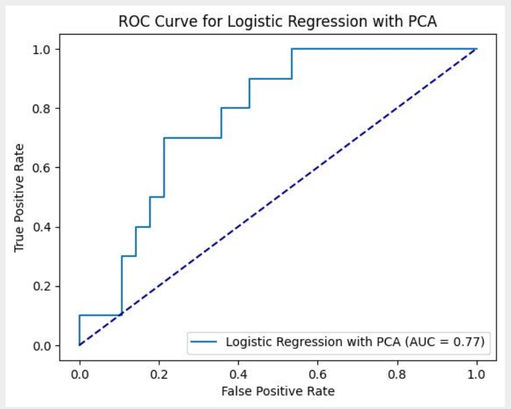


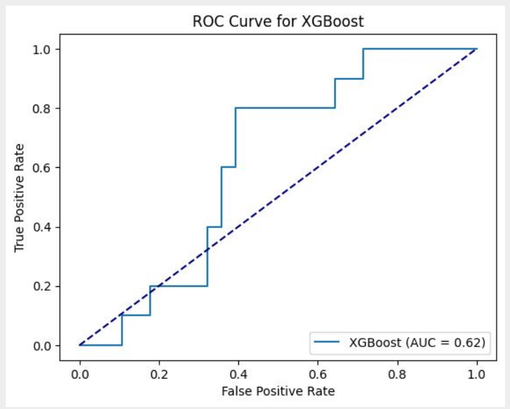


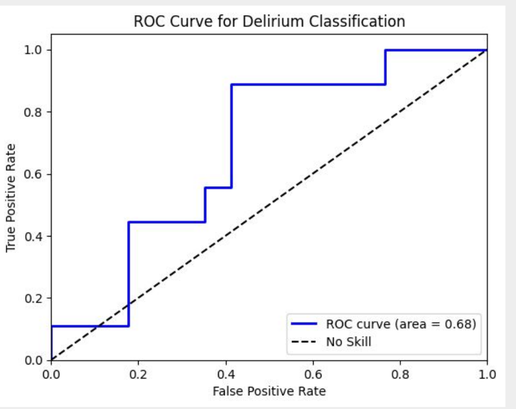






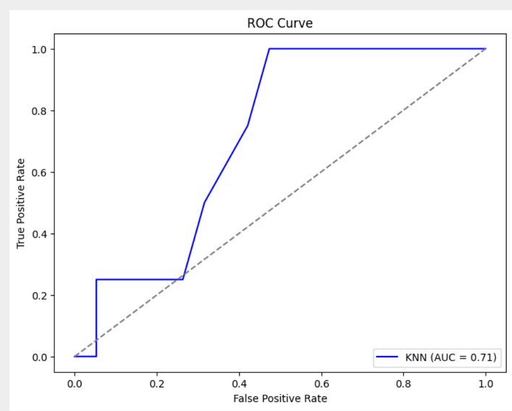


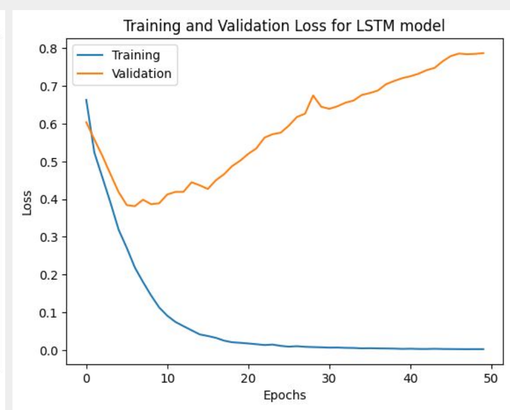


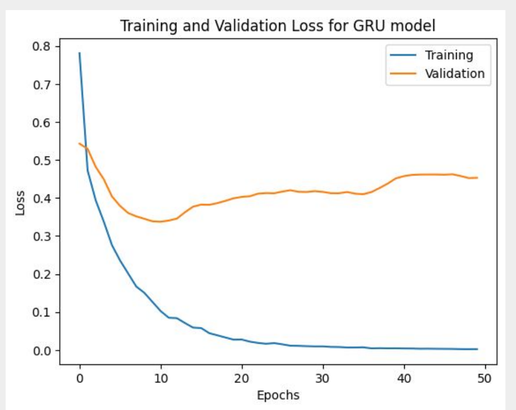


TRANSCRIPTOMIC

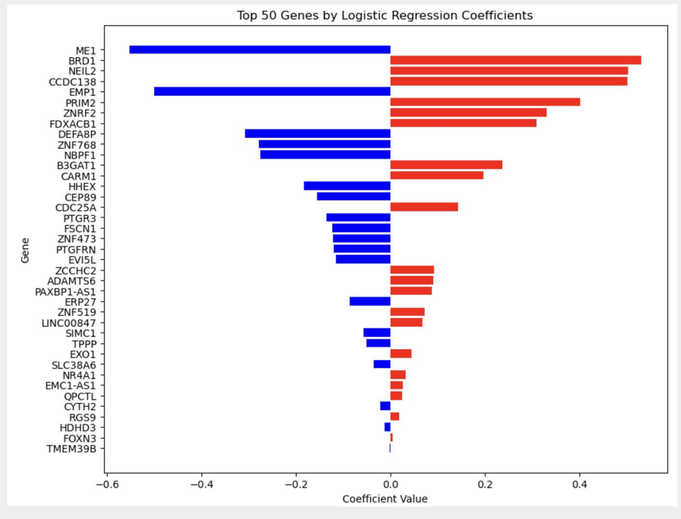
To predict delirium in hospitalized COVID-19 patients, transcriptomic data from peripheral blood mononuclear cells (PBMCs) serially isolated during hospitalization was utilized alongside demographic data. Various machine learning models, including Logistic Regression, Random Forest, Gradient Boost, K-Nearest Neighbors (KNN), and Recurrent Neural Networks (RNN) such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), were employed to identify patterns in gene expression associated with delirium onset. Feature selection for KNN models incorporated Student's t-tests to highlight genes with significant expression changes between Day 4 and Day 7 in delirium and non-delirium cohorts. These selected features were used to train KNN models, achieving an ROC accuracy of 0.71 and a model accuracy of 0.83.







The time-series RNN models used the complete serial gene expression data all days in the dataset, achieving the highest ROC accuracies of 0.86 with LSTM and 0.88 with GRU. Logistic Regression with PCA also showed improved accuracies, further validating the predictive strength of transcriptomic data; its SHAP plot indicating the relative importance of individual genes for the model's prediction. Gene Ontology (GO) Biological Process analysis was conducted on high-importance genes to explore their biological relevance. This approach demonstrates the potential of combining transcriptomic and demographic data to enhance the prediction of delirium, with transcriptomic data providing significantly higher accuracy than models relying solely on demographic features.



CONTRIBUTIONS

Audrey conducted statistical significance testing and causal analysis for predictors of delirium and created data visualizations of the cohort's demographics. Shreya, Grace, Garima, and Matthew created models for admission gene expression and demographic data, with Shreya consolidating all of these models and their analyses. Grace, Matthew, and Garima worked on models for the serial gene expression data; Grace analyzed these results and looked into possible related biological pathways. Garima created and formatted the poster, with help from Matthew and Audrey as well.