Written by Kianna Hendricks
North Carolina A&T State University

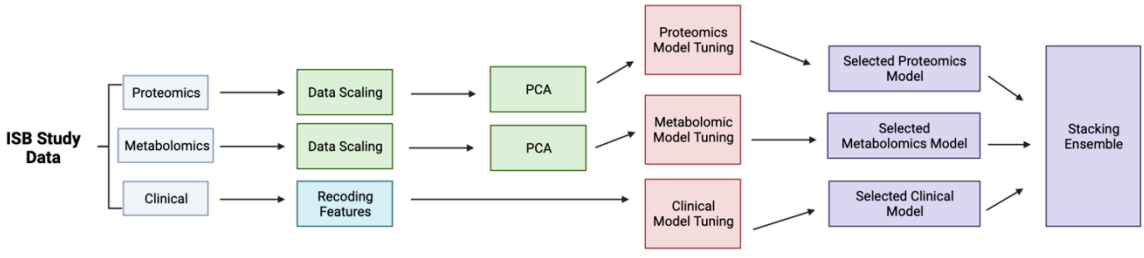
A Stacking Ensemble Approach Leveraging Plasma Proteomics, Metabolomics, and Clinical Data for COVID-19 Severity Prediction

1. Introduction

As of August 13, 2023, there are more than 769 million cases of COVID-19 infections leading to nearly 7 million deaths, showing the immense impact of the pandemic and the long-term challenges the world faces [1]. One of the foremost challenges of COVID-19 management has been the prediction of disease severity. Multi-omics data, particularly plasma proteomics and metabolomics, offers a comprehensive molecular snapshot of an individual's health status. When combined with clinical data, these molecular profiles can provide a more complete understanding of disease severity. Although machine learning (ML) has shown strong predictive capabilities in severity prediction, few studies have integrated these data sources [2–7], and many neglect comprehensive model evaluation metrics and interpretability. Ensemble methods, particularly stacking, show promise in combining the strengths of multi-omics and clinical data. This study proposes a stacking ensemble approach to address these gaps, integrating plasma proteomics, metabolomics, and clinical data with enhanced interpretability for COVID-19 severity prediction.

2. Methodology

Figure 1. The workflow of COVID-19 severity prediction with machine learning



Building upon Dimitsaki et al.'s work [8], this research incorporates plasma metabolomics data and stacking ensemble techniques, offering an enhanced perspective on COVID-19 severity prediction.

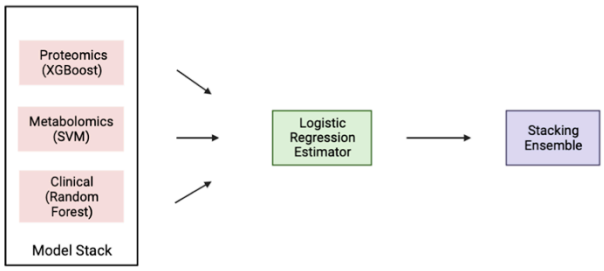
The plasma proteomics, metabolomics, and clinical data were sourced from Su et al. [9], covering a diverse range of patient parameters on a unified set of patients. From 209 initial patients, the dataset, after filtering missing omics data, narrowed to 198 patients.

Table 1: Dimensionality Reduction with PCA

Features	Initial Dimensions	Final Dimensions
Proteins	454	32[1]
Metabolites	891	33[1]

Due to severity class imbalance, the data underwent a 40-20-20 stratified split for train-validation-test sets, ensuring a proportional representation. Clinical data was encoded and 'Unknown' values were imputed. The omics data was standardized and underwent dimensionality reduction via Principal Component Analysis (PCA). Ten classifiers (Logistic Regression, Decision Tree, Random Forest, Extra Trees, Multi-layer Perceptron, Support Vector Machine, AdaBoost, XGBoost, Gradient Boosting, and Quadratic Discriminant Analysis) were initially explored. The models were trained, tested, and validated with a 10-fold stratified cross-validation to maintain balanced class distribution. In model selection, the 'Recall Severe' metric, which measures the recall of the severe class, was prioritized among metrics, such as AUC and F1-score. Consequently, XGBoost was selected for plasma proteomics, Support Vector Machine (SVM) for plasma metabolomics, and Random Forest for clinical data. The goal was to capture diverse patterns to enhance the ensemble's robustness while choosing the highest performing models.

Figure 2. Workflow of the Stacking Ensemble Model Creation



The stacking ensemble was identified as the best ensemble approach. The initial predictions from the base models (SVM, XBG, and Random Forest) are used as input features for the final estimator, a logistic regression. The performance of this ensemble approach was evaluated across various metrics, including accuracy, AUC, macro and weight F1-scores, precision, recall, and, specifically, the recall for the 'severe' class on both test and validation sets. To aid in this interpretability, a method adapted from Dimitsaki et al. [8] was implemented utilizing SHAP values in combination with PCA loadings. This offers an estimated perspective on plasma proteins' and metabolites' feature importance and the need for transparent decision-making, considering model interpretability is important.

3. Results

All performance metrics are reported from the validation set. Table 2 shows the model results for the base models (proteomics, metabolomics, and clinical data) and the stacking ensemble.

Table 2. Performance Metrics of Base Models and Stacking Ensemble Model

Models	Accuracy	AUC	F1-score (macro)	F1-score (weighted)	Precision	Recall Severe
Proteomics (XGB)	0.606	0.819	0.558	0.607	0.553	0.545
Metabolomics (SVM)	0.823	0.913	0.832	0.821	0.813	1.000
Clinical (Random Forest)	0.987	1.000	0.990	0.987	0.990	1.000
Stacking Ensemble	**0.899**	**0.988**	**0.922**	**0.899**	**0.923**	**1.000**

The proteomics model (XGB) has an AUC of 0.819, but an accuracy of 0.606, emphasizing the need for further model refinement. The metabolomics model (SVM) has a perfect 'Recall Severe' score of 1.0, indicating its efficiency in identifying severe COVID-19 cases using metabolomics factors.

Complemented by the robust AUC of 0.913, this model presents reliability and specificity. The clinical model (Random Forest) has nearly perfect scores across metrics. The ensemble method offers robustness with an AUC of 0.988 and a perfect 'Recall Severe' score. Figures 3 and 4 also represent the most influential features derived from the base omics models. Consider the IL6 gene and its related gene entries (P05231.1, P05231.2), which have been previously researched as influential in COVID-19 severity[10–11].

Figures 3. Top influential features from plasma proteomics model

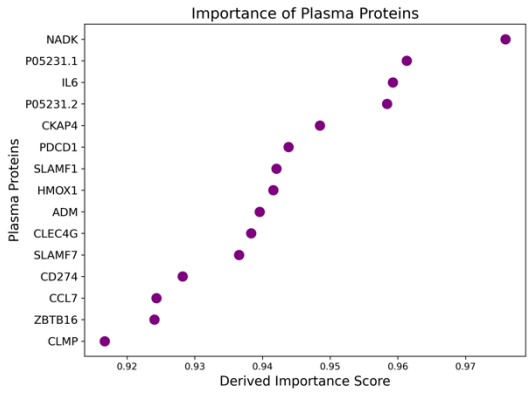
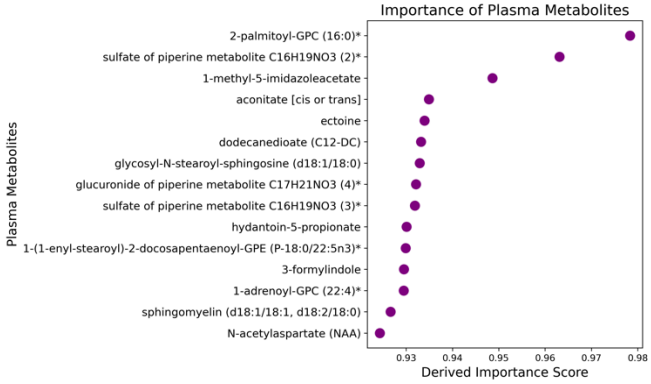


Figure 4. Top influential features from plasma metabolomics model



4. Discussion

Integrating plasma proteomics, metabolomics, and clinical data, the stacking ensemble approach presents promising utility for COVID-19 severity predictions. Blending the base models' strengths, has

ensured a more comprehensive representation of severity predictors. The ensemble has also shown overall improved performance across metrics. The variance in the performance of the plasma proteomics, metabolomics, and clinical models highlights the importance of each data type. While the clinical model showed near-perfect performance all around, integrating multi-omics provided a deeper, and more realistic biological representation of COVID-19 severity. The emphasis on higher performance of the 'Recall Severe' metric ensured the ensemble's ability to minimize false negatives for severe cases. This approach is particularly crucial in healthcare, as misclassifying a severe patient could have dire consequences. Additionally, integrating SHAP values with PCA loading ensures transparency in the base models' predictions. This is another crucial aspect, given the need for explainability for patients and clinicians.

One limitation of this study is the sample size and the class imbalance of severity. Therefore, future research could benefit from larger datasets. Furthermore, while the model prioritizes the recall of severe cases, the potential trade-offs in other metrics like accuracy should be further explored. Finally, the biological implications of the top influential features warrant deeper exploration.

5. Conclusion

This study aimed to implement a stacking ensemble approach incorporating plasma proteomics, metabolomics, and clinical data to predict COVID-19 severity. Based on the performance metrics, this research has shown the potential of plasma multi-omics and clinical data for accurately and reliably predicting COVID-19. As the world grapples with the long-term challenges posed by COVID-19, predictive tools with accuracy and transparency will play a pivotal role in shaping the future of healthcare and public health strategies.

References

[1] "WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data." https://covid19.who.int/table (accessed Aug. 13, 2023).

[2] M. El-Shabrawy *et al.*, "Interleukin-6 and C-reactive protein/albumin ratio as predictors of COVID-19 severity and mortality," *Egypt. J. Bronchol.*, vol. 15, no. 1, p. 5, Jan. 2021, doi: 10.1186/s43168-021-00054-1.

[3] N. Alballa and I. Al-Turaiki, "Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review," *Inform. Med. Unlocked*, vol. 24, p. 100564, Jan. 2021, doi: 10.1016/j.imu.2021.100564.

[4] J. Baj *et al.*, "COVID-19: Specific and Non-Specific Clinical Manifestations and Symptoms: The Current State of Knowledge," *J. Clin. Med.*, vol. 9, no. 6, p. 1753, Jun. 2020, doi: 10.3390/jcm9061753.

[5] J. Zhang, X. Dong, G. Liu, and Y. Gao, "Risk and Protective Factors for COVID-19 Morbidity, Severity, and Mortality," *Clin. Rev. Allergy Immunol.*, vol. 64, no. 1, pp. 90–107, Feb. 2023, doi: 10.1007/s12016-022-08921-5.

[6] B. Shen *et al.*, "Proteomic and Metabolomic Characterization of COVID-19 Patient Sera," *Cell*, vol. 182, no. 1, pp. 59-72.e15, Jul. 2020, doi: 10.1016/j.cell.2020.05.032.

[7] V. R. Richard *et al.*, "Early Prediction of COVID-19 Patient Survival by Targeted Plasma Multi-Omics and Machine Learning," *Mol. Cell. Proteomics*, vol. 21, no. 10, Oct. 2022, doi: 10.1016/j.mcpro.2022.100277.

[8] S. Dimitsaki, G. I. Gavriilidis, V. K. Dimitriadis, and P. Natsiavas, "Benchmarking of Machine Learning classifiers on plasma proteomic for COVID-19 severity prediction through interpretable artificial intelligence," *Artif. Intell. Med.*, vol. 137, p. 102490, Mar. 2023, doi: 10.1016/j.artmed.2023.102490.

[9] Y. Su *et al.*, "Multiple early factors anticipate post-acute COVID-19 sequelae," *Cell*, vol. 185, no. 5, pp. 881-895.e20, Mar. 2022, doi: 10.1016/j.cell.2022.01.014.

[10] P. Sabaka *et al.*, "Role of interleukin 6 as a predictive factor for a severe course of Covid-19: retrospective data analysis of patients from a long-term care facility during Covid-19 outbreak," *BMC Infect. Dis.*, vol. 21, no. 1, p. 308, Mar. 2021, doi: 10.1186/s12879-021-05945-8.

[11] A. Santa Cruz *et al.*, "Interleukin-6 Is a Biomarker for the Development of Fatal Severe Acute Respiratory Syndrome Coronavirus 2 Pneumonia," *Front. Immunol.*, vol. 12, 2021, Accessed: Aug. 14, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2021.613422

[12] P. Yang, Y. H. Yang, B. B. Zhou, and A. Y. Zomaya, "A Review of Ensemble Methods in Bioinformatics," *Curr. Bioinforma.*, vol. 5, no. 4, pp. 296–308.

Appendix

Supplementary Code

For technical specifics, the code is publicly available on GitHub:

https://github.com/kiannaamaya/covid-severity-data-stacked-ensemble/