DS 3001 Final Paper

Kaitlyn Chou [1] **Mackenzie Chen** [2] **Mrunal Kute** [3]

we all worked on coming up with this question

Abstract

Mackenzie wrote the abstract

This research paper examines the following question: do race, underlying medical condition, and age group affect the likelihood of hospitalization status, intensive care unit (ICU) status, or death status due to Coronavirus or COVID-19? Specifically, we wanted to see if a certain race, underlying medical condition, and age group would either increase or decrease the likelihood of hospitalization, ICU, or death status due to COVID-19. COVID-19 was a devastating pandemic that wreaked havoc on the country, so we examined how certain demographic characteristics could disproportionately affect the rate at which a community might experience higher rates of COVID-19. Our research could inform future public health policy by providing insight into what key issues policymakers should focus on when trying to create health programs or guidelines for their community members. Using the *COVID-19 Case Surveillance Public Use Data* from the Centers of Disease Control and Prevention (CDC), we extracted key features, such as race, age, and underlying medical condition in order to generate three models that would predict hospitalization, ICU, and death status from COVID-19. This was appropriate to answer our research question because it included all of the variables we intended to study. To generate this model, we employed supervised learning techniques comparing the performance of multiple algorithms, including logistic regression, single tree models, and random forest, to determine the optimal performance framework. Model training utilized stratified cross-validation to address class imbalances, ensuring robust performance across diverse patient subgroups. Additionally, we optimized model performance by tuning hyperparameters and adjusting classification thresholds to balance sensitivity and specificity. Predictive accuracy was assessed using multiple metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC), providing a comprehensive evaluation of model capability. From our results, we determined that the presence of an underlying medical condition is the most important indicator of whether or not someone will be hospitalized, committed to the ICU, or pass away due to COVID-19. The next most important indicator was age. While race was shown to contribute to the model, it was found to be lower in predictive value in comparison. Overall, our model performance can be analyzed by comparing our ROCs across the three models we used to analyze hospitalization status, ICU status, and death status. Based on our comparison, our AUC interpretations were all pretty excellent, meaning that our models had accurate predictions most of the time.

we all edited the abstract

kaitlyn created the models / did most of the coding for this

1. Data

Mackenzie wrote the data section

Mrunal found the dataset

For this project, we analyzed the *COVID-19 Case Surveillance Public Use Data* from the CDC in order to create a machine learning model to predict the trends of hospitalization, ICU, and death status due to COVID-19 cases in the United States (US) [2]. Specifically, our goal or research question was to see whether there were general correlations between race, underlying medical condition, and age on the probability of a patient's being hospitalized, committed to the ICU, or passing away due to COVID-19. This specific data set was created on May 15th, 2024, and was last updated on July 9th, 2024. Although we had hoped to find more recent data, we were unable to find any data sets with data from the year 2025. We suspect this is due to the fact that the year 2025 is still ongoing, and therefore, datasets haven't been added yet for this year (Figure 18, see Appendix).

The database includes information from U.S. states, territories, affiliates, and other reporting entities. The data itself was appropriate to answer our research question, as it included the necessary variables from millions of patients who had developed COVID-19 since the start of the pandemic. Most of the individuals in our dataset were White/non-Hispanic, with the second largest group being Hispanic/Latino individuals (Figure 21, see Appendix). The age distribution was fairly even across all age groups, and being hospitalized or having an underlying medical condition was pretty common (Figure 22, see Appendix).

kaitlyn helped edit

Based on the data at our disposal, we could accurately and effectively execute our project because we had all the information we needed to analyze trends between the variables. Our research question itself was also very specific and more narrow in scope, so we could easily use our dataset to help answer our question and avoid any challenges, such as overfitting.

The key variables we examined in this dataset included:

- **Race/Ethnicity (`race_ethnicity_combined`)** - The racial background for each case
- **Underlying Health Condition (`medcond_yn`)** - Whether there is an underlying condition present for the case
- **Age Group (`age_group`)** - The age group of the specific case
- **Hospitalization Status (`hosp_yn`)** - Was the case hospitalized?
- **ICU Status (`icu_yn`)** - Was the case admitted to the ICU?
- **Death Status (`death_yn`)** - Did the case pass away?

Working with this dataset presented several challenges. Based on the nature of the dataset and our collective decision to use a classifying decision tree to make the predictions, it was important for us to transform the variables that were "object" data types into "binary" data types. Additionally, since we only needed a small part of the dataset, we had to remove the unnecessary noise and extract our key variables, such as race, age group, underlying medical condition, hospitalization status, ICU status, and death status. The dataset also had a lot of missing or unknown values, which we would have to remove before we started our analysis.

Another major challenge that we had to tackle within our research project was the fact that our previous research question was too large in scope. Not only did it take a really long time to import the dataset with all of the variables, we also had issues with overfitting or producing overly complex models with our classification tree. Specifically, our original research question sought to predict how certain variables, like age group, race/ethnicity, and underlying medical condition, could predict future trends in hospitalization, ICU status, and death due to COVID-19. We quickly realized that our model performance was being affected by excessive noise from irrelevant columns, causing our model to memorize the noise within a large dataset rather than make any substantial future predictions.

Our dataset, in general, is also quite big. It was quite difficult to export the file because it took a really long time. There was also a missing or an under reporting of data because some people might not report their case or follow up. Additionally, we dropped some irrelevant variables that did not contribute to our predictive analysis: the date related to the illness or specimen collection, the date the case was first reported to the CDC, the date of the first positive specimen collection, the symptom onset date, the current status of the patient, and the sex of the patient.

In order to clean the data, we had to drop any null and unknown values, especially cases that have null/unknown values for hospitalization, death, or ICU status - chances are that these might be cases that were not followed up on. We did not have any outliers given the nature of our data. In addition, a lot of the "status" variables are an object data type, which we decided to change to a binary data type so we could use (0, 1) values to represent each case. This made it much easier when reporting statistics on the data. To make this transformation, we one-hot encoded race, age, underlying medical condition, hospitalization status, ICU status, and death status. This way, our algorithm could understand and process our data now that it was in a numerical format. Once we made this conversion, we could put our newly converted variables into a data frame. Finally, we renamed the variables to indicate their binary data type to make it easier to read in our data frame.

Likewise, for more nuanced variables like "Race/Ethnicity", we recoded the categories and standardized them, especially in cases where multiple ethnicities are included. For situations where "unknown" is marked for an ethnicity, that data was ultimately removed. Further, we checked for "impossible data", any data that might have been mis-entered; for example, a negative age or a future date would indicate that we should drop the case. After cleaning the data, multiple sanity checks and accuracy tests were conducted to evaluate the performance of our models. We used visualizations to map out the performance of each model and how they compared to each other.

2. Methods and Results

The analysis began with the data cleaning and preparation outlined above. We extracted key variables required for modeling: race, age, underlying medical condition, hospitalization status, ICU status, and death status. After identifying the exact variable names within the dataset, the data underwent a cleaning process in which missing and "unknown"

values were removed. Because the selected modeling approach relied on classification decision trees, all object-type variables were transformed through one-hot encoding. Race, age, underlying medical condition, and each of the three outcome variables were encoded and compiled into a unified numerical dataframe suitable for machine-learning algorithms.

Before starting to work on our current model, we decided to use a baseline model to have a point of comparison. Logistic regression was selected as the baseline model because it is standard in clinical risk prediction and offers clear interpretability, allowing demographic and medical variables to be linked directly to each outcome. To assess whether nonlinear effects were important, a decision tree classifier was added; this model can capture interactions among predictors and underlying conditions that a linear model cannot. However, because single trees tend to overfit, the analysis emphasized a random forest classifier, which stabilizes predictions through bootstrap aggregation and typically performs well on tabular, imbalanced health datasets. More complex approaches such as support vector machines, gradient-boosted trees, and neural networks were not pursued due to their lower interpretability and limited added value given the structure of the available clinical variables. Overall, this model set balances interpretability with methodological rigor and predictive reliability.

To establish a clear comparison standard, the models were trained using a train–test split with age, race, and underlying medical conditions as predictors, and hospitalization, ICU admission, or death as the target outcome (depending on the model). As stated previously, logistic regression served as the baseline reference. A decision tree was then used to test potential gains from nonlinear modeling, but compared to the baseline, its improvements were small and accompanied by overfitting. The random forest, by contrast, offered consistently stronger and more stable performance, motivating more extensive tuning and threshold optimization for each outcome.

For the hospitalization model, hyperparameters tuned included the number of estimators, minimum samples per leaf, maximum tree depth, and minimum samples required for a split. The parameter class_weight="balanced" was applied to correct for the substantial class imbalance by increasing the weight of the minority class. A precision–recall curve (Figure 1) was used to determine the optimal classification threshold, resulting in a threshold of 0.78; adjusting the threshold allowed finer control over the trade-off between precision and recall (Figure 1). StratifiedKFold cross-validation was employed to estimate model stability, as this method preserves class proportions in each fold, a critical feature for imbalanced datasets. Cross-validation results indicated that the model generalized effectively.

The models were evaluated using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC), as these metrics capture complementary aspects of performance under the extreme class imbalance present in the CDC COVID-19 surveillance data. Because the vast majority of individuals in the dataset were not hospitalized, not admitted to the ICU, and did not die, accuracy alone is misleading: a classifier can achieve high accuracy by predicting only the majority class while failing to identify the far smaller proportion of severe cases. Precision and recall therefore play a more critical role, as they quantify how often predicted severe cases are correct and how effectively true hospitalization, ICU, and death cases are detected. This is particularly important in this dataset, where models frequently exhibited very low recall for hospitalization and ICU outcomes due to the rarity of positive cases. The F1-score provides a balanced measure that penalizes this imbalance between precision and recall, making it more informative than accuracy for these outcomes. AUC further evaluates each model's ability to distinguish between severe and non-severe cases across all possible thresholds, which is essential in public health applications where threshold sensitivity directly affects risk classification.

To strengthen the robustness of these evaluations, stratified K-fold cross-validation was applied to ensure that each fold preserved the same extremely low proportion of hospitalization, ICU, and death cases found in the full dataset, preventing folds from containing no positive cases at all. Threshold tuning was additionally necessary because the default 0.50 decision threshold consistently resulted in models that predicted almost no positive outcomes, especially for hospitalization and ICU admission. Adjusting thresholds based on precision–recall tradeoffs improved sensitivity and produced more clinically meaningful predictions. Collectively, these evaluation strategies provide a more reliable assessment of performance and increase confidence that the models generalize appropriately given the structure and imbalance of the underlying COVID-19 dataset.

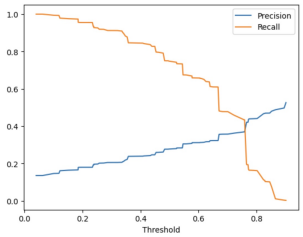

Figure 1. Precision/Recall Threshold Curve for Hospitalization Model

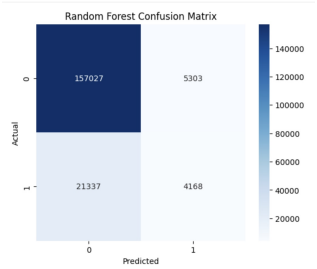

Figure 2. Hospitalization Random Forest Confusion Matrix

The ICU model underwent the same hyperparameter adjustments, including the application of balanced class weights. A precision–recall curve was again used to select an appropriate classification threshold, and a value of 0.80 was chosen (Figure 3). StratifiedKFold cross-validation demonstrated strong model stability.

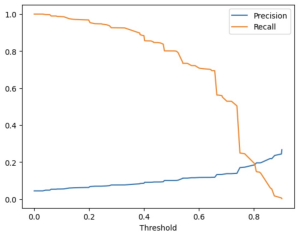

Figure 3. Precision/Recall Threshold Curve for ICU Model

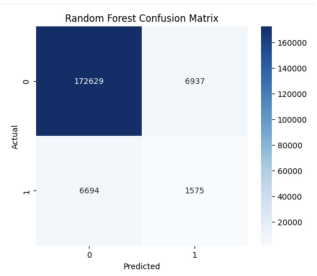

Figure 4. ICU Random Forest Confusion Matrix

The death model incorporated the same core hyperparameters, as well as the parameter max_features="log2," which improved performance relative to the default "sqrt" option. Threshold tuning based on a precision–recall curve produced an optimal threshold of 0.8, however, a threshold of 0.70 was chosen to try and increase the amount of class 1 (death) predictions (Figure 5). StratifiedKFold cross-validation again showed strong generalization.

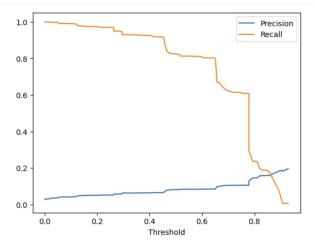

Figure 5. Precision/Recall Threshold Curve for Death Model

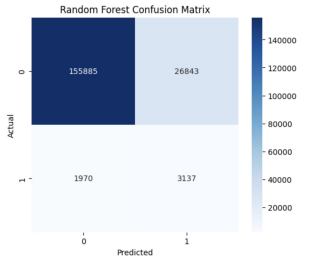

Figure 6. Death Random Forest Confusion Matrix

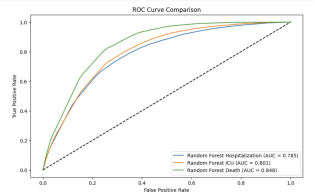

Figure 7. ROC Curve Comparison

To further assess model performance, confusion matrices were analyzed. For example, the confusion matrix for the hospitalization model (Figure 2) shows that the classifier overwhelmingly predicted class 0, correctly identifying true negatives but failing to identify any true positives. False negatives were common, while false positives and true positives were absent. This pattern confirmed the severe class imbalance and underscored the need for class weighting in both decision tree and random forest models. The same imbalance patterns were observed in the ICU (Figure 4), and death models (Figure 6) were addressed through identical weighting strategies.

2.1. Main Results

Model outputs showed that the hospitalization classifier performed strongly for class 0 (non-hospitalized cases) but frequently failed to identify class 1 cases, resulting in a high number of false negatives. Although overall accuracy was 85.8%, this value was inflated by the dominance of non-hospitalized observations (Figure 10). The ROC AUC of

0.785 indicated that the model maintained substantial discriminatory power despite low recall for the positive class. According to the hospitalization classification tree, which is pretty deep, complex, and has many sequential conditions, the Gini impurity decreases with each split and gets closer to 0, becoming more certain of its predictions (Figure 8). Overall, the Gini values are relatively high; this is likely due to a class imbalance. The tree is fairly complex, with multiple splits and nodes; the orange and blue nodes represent two different classes, with darker colors indicating higher confidence and lighter colors indicating lower confidence. The model also captures non-linear interactions, using "if-then" decisions to determine where the next split will be. Feature importance analysis showed that the presence of an underlying medical condition was the strongest predictor of hospitalization, followed by age, with the 60–69 and 10–19 age groups emerging as the most influential (Figure 9).

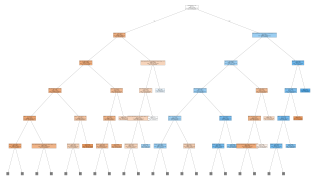

Figure 8. Hospitalization Classification Tree

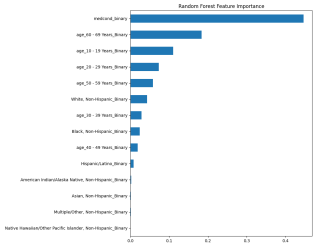

Figure 9. Hospitalization Random Forest Feature Importance

	precision	recall	f1-score	support
0	0.88	0.97	0.92	162330
1	0.44	0.16	0.24	25505
accuracy			0.86	187835
macro avg	0.66	0.57	0.58	187835
weighted avg	0.82	0.86	0.83	187835

Figure 10. Hospitalization Classification Report

As with the hospitalization model, the ICU classifier showed very high performance for class 0 but insufficient sensitivity to class 1, generating many false negatives. Accuracy reached 92.7%, but this reflected the exceptionally skewed class distribution rather than genuine predictive accuracy for ICU admissions (Figure 13). The ROC AUC of 0.801 indicated moderate discriminatory capability. Our ICU classification tree was shown to be relatively similar to our hospitalization classification tree, with relatively high Gini impurity values indicating a class imbalance. Multiple nodes and splits suggest the complexity of our tree and slight overfitting, which we tried to account for by using the random forest classifier and adjusting hyperparameters (Figure 11). Feature importance rankings were consistent with those of the hospitalization model, with underlying medical condition serving as the strongest predictor, followed by age—particularly the 60–69 and 10–19 groups (Figure 12).

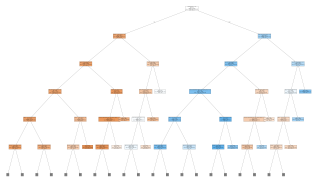

Figure 11. ICU Classification Tree

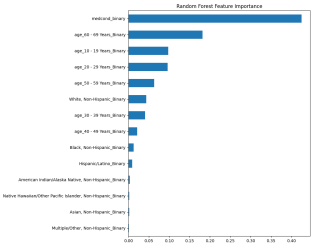

Figure 12. ICU Random Forest Feature Importance

	precision	recall	f1-score	support
0	0.96	0.96	0.96	179566
1	0.19	0.19	0.19	8269
accuracy			0.93	187835
macro avg	0.57	0.58	0.57	187835
weighted avg	0.93	0.93	0.93	187835

Figure 13. ICU Classification Report

Unlike the hospitalization and ICU models, the death model demonstrated high recall for the positive class and successfully identified most death cases (Figure 16). However, this improvement came at the cost of low precision, producing a substantial number of false positives. The model achieved an accuracy of 84.7% and a ROC AUC of 0.848, indicating strong class separability. Precision and recall for class 0 were high (0.99 and 0.85, respectively), whereas precision for class 1 remained low due to the elevated false-positive rate. Once again, our classification tree was complex, with its predictions showcased through different splits at each of the nodes. The same order of predictors appeared, with underlying medical condition being the most predictor (Figure 14). Examination of feature importance revealed a similar hierarchy: underlying medical condition was the dominant

predictor, followed by age groups 60–69 and 20–29 (Figure 15).

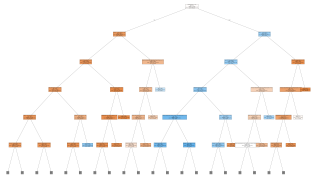

Figure 14. Death Classification Tree

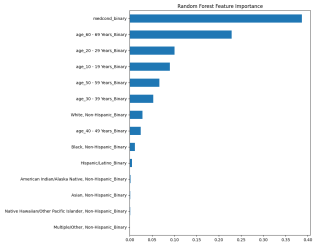

Figure 15. Death Random Forest Feature Importance

Kaitlyn generated the models (reports 3 out of them)

Classification Report:	precision	recall	f1-score	support
0	0.99	0.85	0.92	182728
1	0.10	0.61	0.18	5107
accuracy			0.85	187835
macro avg	0.55	0.73	0.55	187835
weighted avg	0.96	0.85	0.90	187835

Figure 16. Death Classification Report

Across all three models, a consistent pattern emerged. Underlying medical condition was the most influential predictor of adverse COVID-19 outcomes, followed by age. Although race contributed to predictions, it played a substantially smaller role compared to medical conditions and age. Among race-related predictors, the distinction between White and non-White individuals appeared most significant. The age groups 60–69 and 10–19 consistently ranked among the top predictors across models.

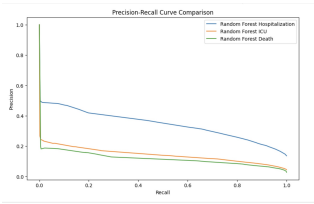

Figure 17. Precision-Recall Curve Comparison

2.2. Confidence in Results

Overall, confidence in the model results is low due to the large mismatch between precision and recall. While an inverse relationship between these metrics is expected, clinical models such as these require a careful balance because lives depend on the accurate identification of both positive and negative outcomes. However, in the case of our particular models, the hospitalization model shows low recall, the death model shows low precision, and ICU model shows low values for both precision and recall. Notably, these issues apply specifically to class 1 predictions, which indicate whether a case is positive; poor predictions on whether a case is positive mean a model is not very reliable.

However, the ROC AUC values are reasonably high across all models, suggesting that the underlying ability to separate positive from negative cases is acceptable. The core problems appear to stem from threshold selection and handling of class imbalance. With additional tuning such as threshold adjustment, resampling, or class weighting, the models could likely become more reliable, which would improve overall confidence in the results.

3. Conclusion

Jamaal wrote this section

This project examined the effects of race, age group, and underlying medical conditions on the likelihood of hospitalization, ICU admission, or death among individuals diagnosed with COVID-19 in the United States. Using the CDC COVID-19 Case Surveillance Public Use dataset and employing a sequence of supervised learning methods, the analysis revealed consistent and meaningful patterns across all three clinical outcomes. The presence of underlying medical conditions emerged as the strongest predictor of severe outcomes, followed by age group, while race contributed only marginally once health status and age were taken into account. These findings provide empirical support for the existing clinical understanding of COVID-19 risk factors and speak to broader implications for public health policies, triage prioritization, resource allocation, and the dissemination of public information.

The results across the hospitalization, ICU, and mortality models consistently pointed to underlying medical conditions as the variable with the highest feature importance. This aligns with findings from other research studies demonstrating that comorbidities such as diabetes, hypertension, and chronic lung disease substantially increase ICU need and mortality [6]. Similarly, the strong predictive role of age, particularly for individuals aged 60–69 and older, reflects well-established age-related vulnerability documented throughout the pandemic. Race, while included as a predictor, contributed much less meaningfully than the clinical variables. This pattern is consistent with research showing that race often correlates with social and structural determinants of health rather than being a biological risk factor itself [4]. In this sense, the limited predictive strength of

race in the model may reflect both the CDC dataset's high proportion of missing race/ethnicity data and the absence of socioeconomic features that typically explain disparities.

There are several valid criticisms that can be made about our models. One key issue is the imbalance between precision and recall, which can reduce the reliability of the predictions. While we acknowledge this limitation, we took deliberate steps to address it. We examined the full precision–recall curve and selected the threshold that provided the strongest overall performance for our objectives. We also accounted for the dataset's class imbalance by applying techniques such as setting class_weight="balanced", ensuring that the minority class was properly represented during training.

A second concern is overfitting, particularly since decision trees are prone to it. To mitigate this, we used a random forest model, performed cross-validation, and tuned both hyperparameters and classification thresholds. A related criticism is that decision trees can be unstable across random seeds. This instability is expected, but the use of a random forest helps counteract it by aggregating predictions across many trees. Although individual tree structures may vary, the ensemble's overall predictions are typically far more stable and reliable.

Although our models achieved respectable ROC AUC scores (0.78–0.85 across outcomes), as mentioned previously, the inherent class imbalance in the dataset significantly affected predictive performance. The hospitalization and ICU models exhibited low recall for the minority class, leading to high false-negative rates, whereas the death model produced many false positives despite strong recall. These challenges highlight a central limitation of applying decision-tree–based models to highly imbalanced clinical data. Although class weighting and stratified cross-validation improved stability, more advanced imbalance-handling strategies, such as SMOTE or cost-sensitive learning, could have further improved minority-class detection. For example, synthetic oversampling methods could have strengthened decision boundaries by creating additional realistic samples for severe outcomes, a strategy shown in prior work to boost recall in COVID-19 severity prediction [1]. Additionally, the CDC dataset's missing values and coarse variable groupings constrained both model granularity and inference strength.

Another limitation is the reliance on relatively simple supervised learning methods. Although logistic regression, decision trees, and random forests offer interpretability and computational efficiency, they are not the most powerful techniques for risk prediction in clinical settings. Several studies on COVID-19 outcome prediction, such as the COVID-GRAM risk score and the ISARIC WHO Clinical Characterization Protocol models have relied on penalized regression and ensemble stacking methods to achieve higher discriminatory performance. These models also benefit from richer clinical variables that can provide specific details; One study conducted that used these methods focused on variables such as vital signs, comorbidity counts, laboratory markers, and imaging studies, which were clearly unavailable in the public CDC dataset [3]. As a result, the current model's present predictive capabilities are limited by both data richness and model sophistication.

Despite these constraints, this project demonstrates that even with simple public-use data, it is possible to extract meaningful insights about the clinical variables most strongly associated with severe COVID-19 outcomes. The consistent ordering of predictor importance across all three outcomes increases confidence in the generalizations made in the findings. Moreover, the successful application of multiple evaluation metrics, including ROC AUC, precision-recall analysis, and confusion matrices, provides a more nuanced understanding of the strengths and weaknesses of each model.

Additionally, future work could extend this analysis in several directions. First, the inclusion of more sophisticated machine learning methods, such as Gradient Boosted Trees (XGBoost, LightGBM, CatBoost) or regularized logistic regression, could improve sensitivity to minority classes while reducing overfitting. Second, implementing oversampling or synthetic minority data generation, such as SMOTE, would help address the severe class imbalance that limited the performance of the current models. Third, incorporating more socioeconomic variables like income, education, vaccination status, and longitudinal markers of health could dramatically expand the explanatory power of the model. Fourth, interpretation methods such as SHAP values would provide more granular insights into variable interactions and individualized risk patterns. Finally, expanding the modeling framework to include survival analysis could capture time-to-event patterns, an approach commonly used in clinical risk modeling but not used here due to data constraints.

In conclusion, the project aimed to successfully identify the primary predictors of severe COVID-19 outcomes and evaluate multiple modeling strategies using a large-scale national dataset. Although the limitations of class imbalance, missingness, and granularity constrained the model performance, the analysis still provides a very rigorous foundation for understanding how demographic and clinical factors interact to shape patient risk. The model's findings emphasize the importance of comorbidity screening and age-adjusted triage protocols and point toward several promising directions for future research that could enhance predictive accuracy and public health utility.

4. Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

5. References

1. Alizadehsani, R., Alizadeh Sani, Z., Behjati, M., Roshanzamir, M., Hussain, S., Abedini, N., … Nahavandi, S. (2021). Risk factor prediction, clinical outcomes, and mortality in COVID-19 patients using machine learning models. *Computers in Biology and Medicine, 132*, 104304. https://doi.org/10.1016/j.compbiomed.2021.104304

2. Centers for Disease Control and Prevention. (n.d.). *COVID-19 Case Surveillance Public Use Data* [Data set]. Data.CDC.gov. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

3. Liang, W., Liang, H., Ou, L., Chen, B., Chen, A., Li, C., Li, Y., … He, J. (2020). Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Internal Medicine, 180*(8), 1081–1089. https://doi.org/10.1001/jamainternmed.2020.2033

4. Mackey, K., Ayers, C. C., Kondo, K. K., Saha, S., Advani, S. M., Young, S., Spencer, H., Ralls, M. W., Anderson, J., Veazie, S., Kansagara, D. (2021). Racial and ethnic disparities in COVID-19–related infections, hospitalizations, and deaths: A systematic review. *Annals of Internal Medicine, 174*(3), 362–373. https://doi.org/10.7326/M20-6306

5. OpenAI. (2025, December 10). ChatGPT [GPT-5]. https://chat.openai.com/

6. Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C., Curtis, H. J., Mehrkar, A., Evans, D., Inglesby, P., Cockburn, J., McDonald, H., MacKenna, B., Tomlinson, L., Douglas, I. J., Rentsch, C. T., Mathur, R., Wong, A. Y. S., Grieve, R., & Goldacre, B. (2020). Factors associated with COVID-19–related death using OpenSAFELY. *Nature, 584*, 430–436. https://doi.org/10.1038/s41586-020-2521-4

6. Appendix

Item	Value
Dataset Source	CDC COVID-19 Case Surveillance Public Use Data
Number of Observations (after cleaning)	939, 174
Variables Used	Race, Age Group, Underlying Medical Condition, Hospitalization Status, ICU Status, Death Status
Date Created	May 15, 2024
Last Updated	July 9, 2024

Figure 18. Data Summary Table

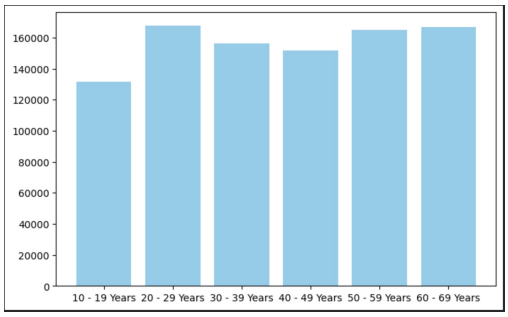

Figure 19. Age Histogram

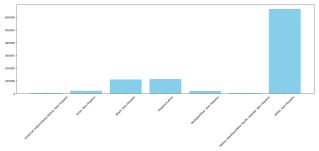

Figure 20. Race Histogram

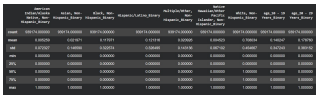

Figure 21. Basic Table Statistics (Part 1)

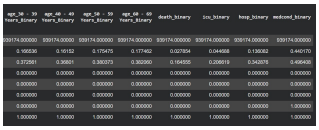

Figure 22. Basic Table Statistics (Part 2)