COVID-19 Severity and Mortality Analysis

Group Members: *Saithanusri Avirneni (savirneni3), Yoonha Kim (ykim3056), Harsha Ramesh (hramesh8), Maria Tariq (mtariq37), Thao Truong (ttruong42)*

1.0 Introduction and Problem Definition

1.1 Objective

Our goal is to create a predictive model for assessing COVID severity and mortality risk using the Kaggle Mexican patient COVID-19 dataset [1]. COVID severity and mortality will be evaluated by examining patient records to determine the extent of medical interventions required, such as hospitalization, ICU admission, intubation, and mortality. This investigation will also explore whether specific demographics or underlying health conditions correlate with a higher risk of severe COVID outcomes.

Model development in Python will involve exploring multiple models to select the most accurate one. Tableau will be used for an interactive visualization, providing insights and enabling real-time outcome simulations for patients. Our approach ensures a comprehensive understanding of the influence of patient data, including health conditions, on COVID outcomes.

1.2 Current Methodology & Limitations

Cascella [2] states COVID as the cause of over 6 million deaths in 2020. Despite vaccines, persistent outbreaks due to mutant variants call for real-time diagnostics. Current severity assessment relies on symptom monitoring, health analysis, clinical evaluations, and PCR tests. Berlin's [3] research stresses prompt treatment for severe COVID to avoid negative outcomes. However, limitations arise from Cascella's lack of live predictions and Berlin's focus on only pulmonary complications. Our work addresses these gaps, innovatively assessing a wider range of conditions and patient information and providing live outputs.

2.0 Literature Survey

Bonand[4], Romero[5], and Hu[6] conducted a comprehensive analysis of COVID with a primary focus on age, establishing its significant role in predicting mortality. Romero's work found that the risk of mortality increased by 7.4% for every age year. Although their research acknowledges age will exacerbate other health conditions, it does not explicitly delve into the interplay between age and other influencing factors. Based on their research, our model will also include age, but will improve on their work by incorporating additional patient health data.

Kompaniyets[7] and Treskova[8], Bloom [9] studies explore the link between pre-existing health conditions and severe COVID outcomes. Their work highlights the influence of common underlying health factors such as respiratory conditions, hypertension, obesity, and diabetes on exacerbating COVID severity, which is useful for our model feature selection process. However, our model will expand on their work by considering additional, less common risk factors such as various medical conditions, demographics (including sex, pregnancy, and tobacco usage), and their associations with COVID severity.

Quiroz-Juárez[10] and Casillas[11] developed machine learning models to predict high risk COVID patients. Quiroz-Juarez used only supervised learning models, and Casillas data found the XGB method to be the most effective. Their work provides a foundation for our model selection and guides our approach of testing several models. We will expand on their approach by potentially testing unsupervised models and other boosting methods.

Moulaei's[12] research emphasizes data balancing, evaluating model performance, and using a confusion matrix to reveal false positives and negatives. Chens[13] predictive model evaluated effectiveness of identifying high risk patients through metrics like accuracy, sensitivity and AUC-ROC. These methods will be useful for our project to assess and compare the different models. We aim to combine these methods for a more comprehensive assessment of the accuracy and strength of our model.

Gao's[14] study reviews risk factors for severe COVID, the work acknowledges limitations like speculative risk associations and potential underestimations of chronic conditions due to testing hesitancy, emphasizing the need for cautious interpretation. Albitar [15] and Vaid [16] find demographic conditions as significant predictors of death, urging further investigation. Vaid cautioned against region-specific mortality models, highlighting poor generalization. Despite lacking demographic or socioeconomic data, we address this by assuming consistent disease impact on COVID across demographics.

Musheyev[17] focuses on a predictive model for COVID outcomes, emphasizing early high-risk patient identification and personalized post-discharge care. The limitation is the need for continuous updates; our approach improves this with real-time data integration, ensuring ongoing adaptability in patient care.

3.0 Proposed Method

Our model leverages a comprehensive dataset and advanced machine learning to capture complex factors contributing to severe outcomes, allowing our model to surpass the predictive capabilities of traditional diagnostic tools. Real-time predictions enable prompt interventions, potentially saving lives by targeting high-risk individuals effectively. Our approach introduces several innovative features not identified in existing literature: COVID Severity Prediction, Advanced Visualization, Patient Risk Simulation.

Our intuition suggests that our approach surpasses the state of the art by addressing current limitations and introducing innovative features. Current models primarily focus on COVID diagnosis or mortality prediction. We believe that by shifting the focus to predicting COVID severity and incorporating outcomes such as hospitalization, ICU admission, intubation, and mortality, our approach offers a more holistic understanding of the disease progression and patient outcomes. Additionally, our focus on advanced visualizations and real-time simulation within Tableau integrates both descriptive and predictive analytics. These capabilities address critical gaps in existing models, which often only offer descriptive insights. As a result, healthcare professionals will be equipped with actionable insights via a user-friendly interface that can be implemented in healthcare settings to support COVID patient management.

3.1 Algorithm

A Random Forest model was built in Python and used to predict patients at high risk of needing medical interventions based on underlying health data and age. The algorithm work consisted of three main steps: data preparation, model development and testing, model deployment

Data Preparation: New feature columns were generated, including ones for patient death, COVID status, and the sum of comorbidities. Additionally, a binary "AT_RISK" variable was created to identify high-risk COVID patients based on their outcomes (ICU,hospitalization,Intubation,Death). Missing data for health conditions were removed, and columns were converted into binary format to aid modeling.

Model Development: The model was iteratively improved and validated through rigorous testing to ensure a framework capable of accurately predicting COVID risk. To optimize our model we refined the representation of influential features, particularly age into one hot encoded column, and removed 5 non influential columns. Our final model had an accuracy of ~66% and used patient gender, age and 7 health conditions to generate predictions for At Risk patients.

Model Deployment: The final trained random forest model (deployed via TabPY), along with two datasets created during model development, were used to create a complex interactive Tableau dashboard.

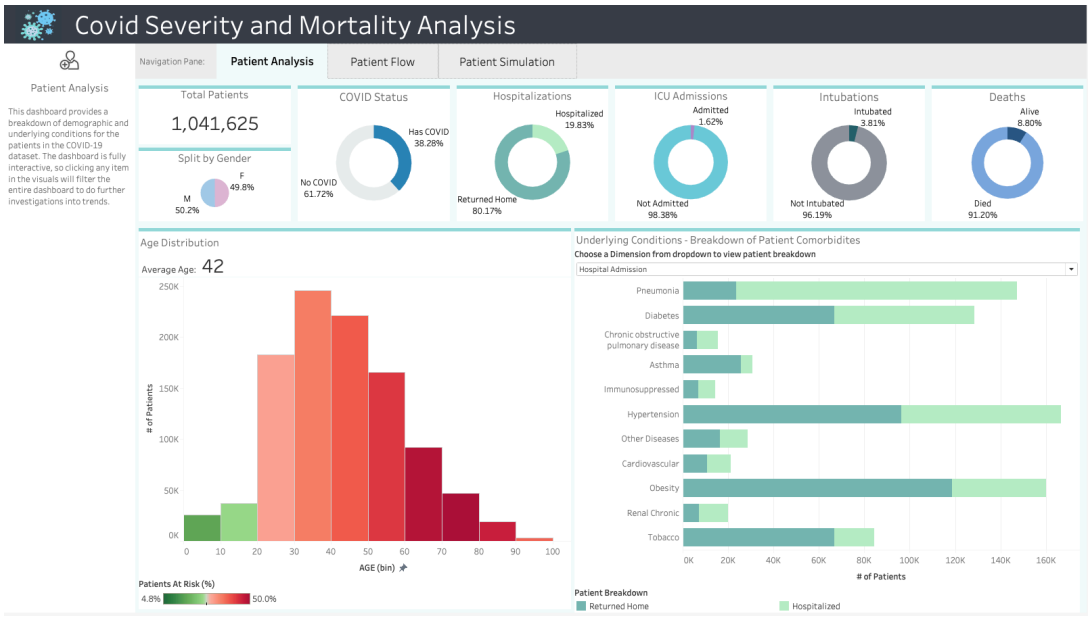
3.2 User Interface Visualization

The team utilized Tableau to develop an intuitive and interactive UI for exploring COVID-19 severity and mortality predictions. Integration with the TabPy extension enabled real-time execution of

.

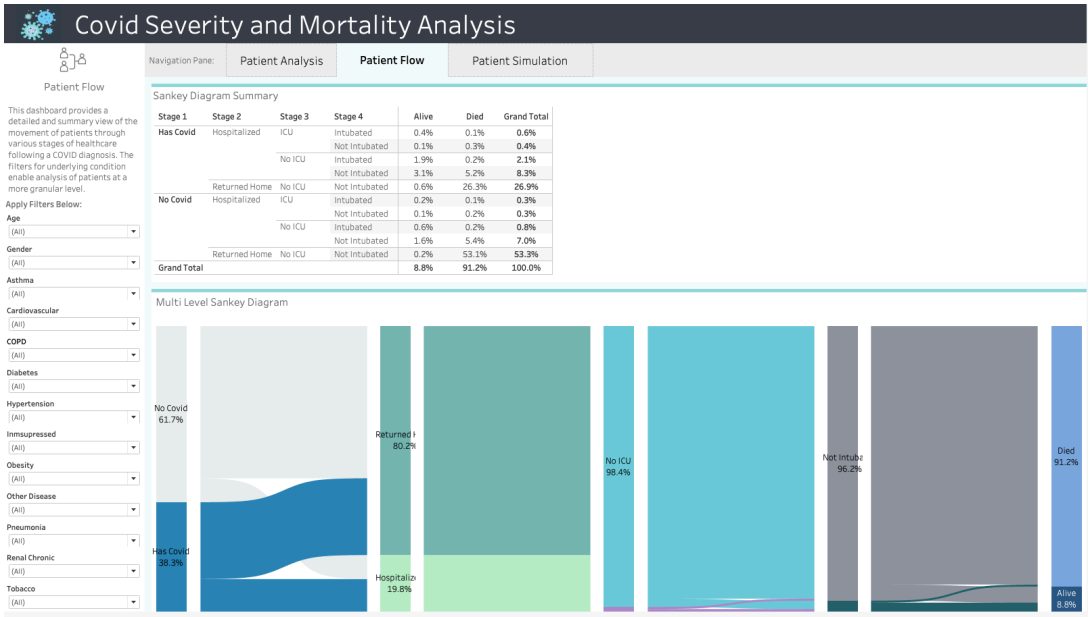
predictive models within Tableau, facilitating dynamic visualization of results. A link to the partial dashboard is available at: [Partial Dashboard Link] However, the full dashboard must be launched locally due to the TabPy Connection, instructions are included.

Patient Analysis



The Patient Analysis dashboard provides comprehensive insights into patient risk factors, presenting high-level values for medical interventions in the data, alongside a dynamic breakdown of underlying conditions. Users can click on visuals to filter the dashboard for deeper trend analysis, facilitating exploration of both overarching patterns and specific details.

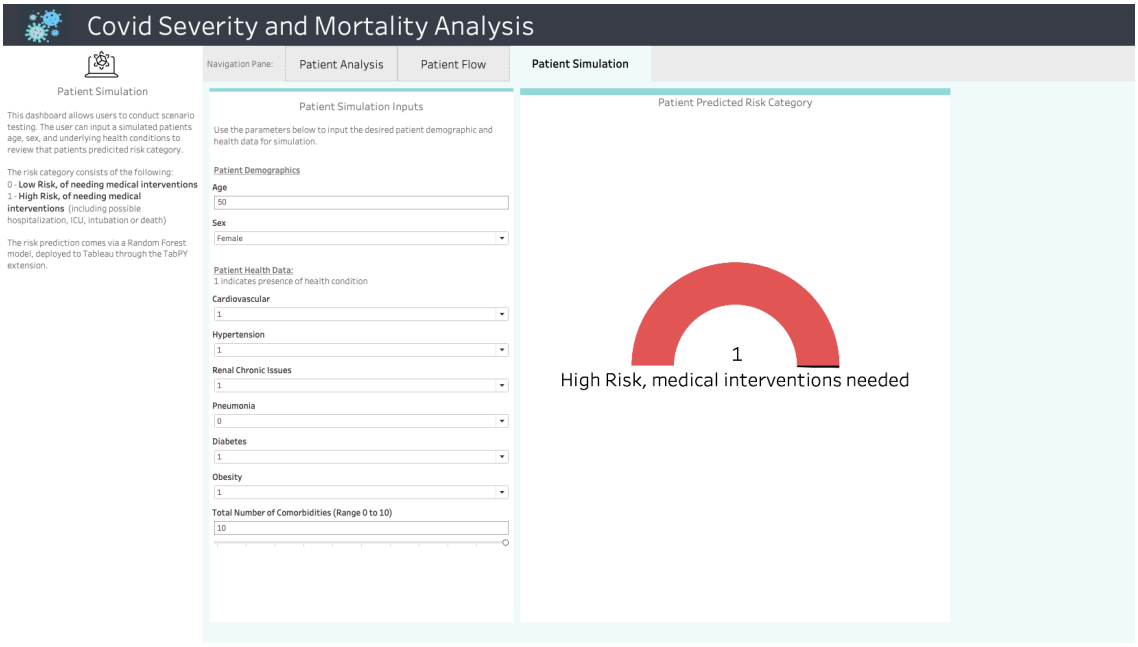
Patient Flow



.

The Patient Flow dashboard provides detailed views of patient movement post-COVID diagnosis, utilizing multilevel Sankey visualizations to illustrate pathways across intervention stages. Users can apply filters for underlying conditions to conduct granular analysis, enriching insights into patient trajectories and outcomes.

Patient Simulation



The Patient Simulation dashboard enables users to conduct scenario testing by creating simulated patients and observing predicted risk classifications. User inputs, including demographics (age and gender) and health conditions, dynamically adjust a gauge chart to display risk prediction. This functionality is driven by the trained Random Forest algorithm seamlessly integrated into Tableau through the TabPY extension. Utilizing Tableau parameters as the inputs for Tableau's script_real calculation, users can effectively tailor simulations to their specific needs, facilitating informed decision-making and risk assessment.

4.0 Experiments and Evaluation

4.1 Random Forest Model Testing

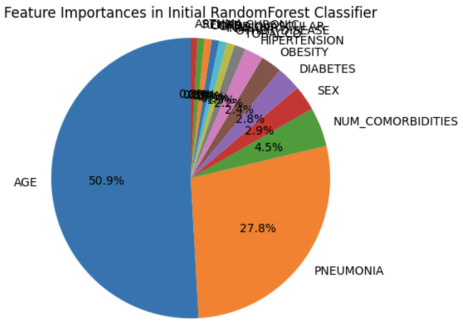
For initial data modeling, we employed the RandomForestClassifier model, using the 'AT_RISK' column as the target, and all dataset columns as features. We defined a patient as at risk if the conditions for any of the columns ICU, Death, and intubation are met when a patient has a COVID positive test. The dataset was partitioned into training (70%) and testing (30%) sets, adhering to the standard practices for model evaluation. The initial model achieved an accuracy of approximately 61.45%.

Our experimental framework was crafted using a randomly selected subset that comprised 50% of the original dataset, chosen to maximize variability and representativeness. The primary goal of our experiments was to identify the features that significantly impact the 'AT_RISK' status and to evaluate the predictive accuracy of our model in pinpointing individuals at high risk. To enhance the reliability of our results, we conducted five iterations of the random first model, each with a different 50% sample of the dataset. The average accuracy of these trials was approximately 61.80%, with a standard deviation of 0.17%. These findings indicate that our model effectively generalized from the patterns learned in the

.

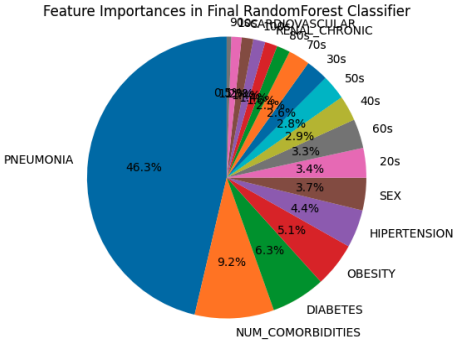
training data to unseen data, with consistent performance across multiple samples. This methodology underscores the thoroughness of our approach, providing robust evidence of the model's capabilities in identifying at-risk individuals with high reliability and accuracy.

4.2 Feature Importance Analysis



The feature importance analysis was vital in identifying key predictors. We initially encoded categorical variables into numerical formats for compatibility with the RandomForest algorithm. This facilitated ranking by importance, revealing 'AGE' and 'PNEUMONIA' as the most influential, with scores of 50.9% and 27.8%, respectively. Other features like 'NUM_COMORBIDITIES', 'SEX', and 'DIABETES', while less impactful, still contributed significantly to the model's decision-making.

4.3 Improve Random Forest Model



In response to our initial analysis using the Random Forest model, we observed a disproportionately high feature importance score for 'AGE', accounting for nearly 50% of the predictive outcome. This indicated a potential overemphasis on age within the model, which could overshadow other significant predictors.

To address this and enhance the model's generalization capabilities, we implemented a strategic modification to the representation of age. Recognizing that age might influence risk factors in a non-linear manner, we categorized age into decades - (i.e'10s', '20s' etc.). This categorization helps to encapsulate the varying effects of age across different life stages, reducing the overwhelming impact of the continuous age variable.

Based on importance scores from our previous model, we streamlined the model by dropping features with minimal impact, like 'INMSUPR', 'ASTHMA', 'TOBACCO', 'COPD', and 'OTHER_DISEASE'. These are now represented in the NUM_COMORBIDITIES column.

These refinements led to a significant improvement in our model's performance. The average Random Forest accuracy increased to 66%, which is a ~ 4% improvement over the previous model. This enhancement not only demonstrates the effectiveness of feature categorization and refinement in predictive modeling but also underscores the importance of continuous model evaluation and optimization

4.4 Multilayer Perceptron (MLP) Model

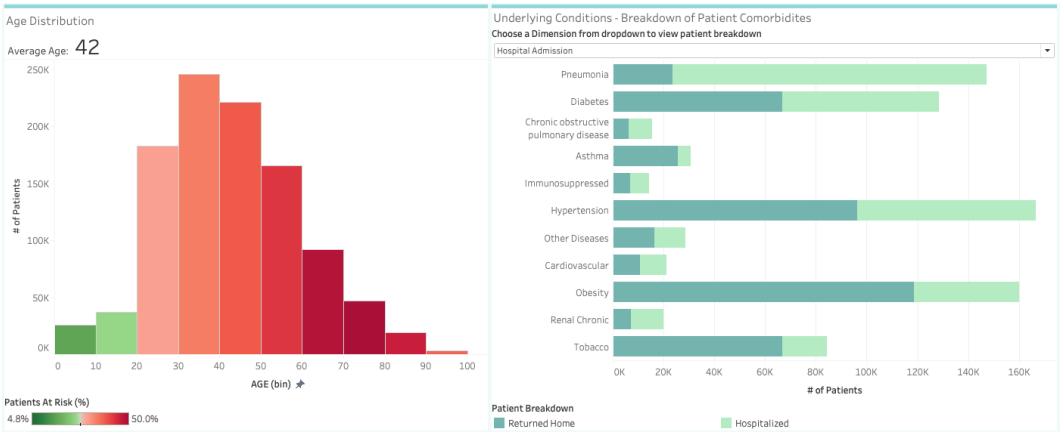
In the final phase, we evaluated the accuracy of our model against the Multi-layer Perceptron (MLP) to ensure alignment with other models. The MLP, is a neural network classifier implemented using scikit-learn's MLPClassifier, which underwent configuration adjustments to capture complex feature interactions. It employed a 'ReLU' activation function and an 'sgd' solver for efficiency with large datasets. With an adaptive learning rate, the model achieved an accuracy of 71% on a 70-30 training-testing dataset split. Despite the MLP's higher accuracy, due to its longer processing time, we ultimately opted to proceed with the random forest. The table below displays the scores of all models:

Model	Accuracy	Precision
Initial Random Forest Model	62%	63%
Improved Random Forest Model	66%	65%
Multilayer Perceptron Model (MLP)	71%	70%

.

4.4 Patient Health Condition Impacts

Our dashboard visuals further show that age and Pneumonia have a high impact on patient risk, further confirming what we see in the data. From the age histogram on the left we can see that risk percentage gets progressively higher as age increases, and the bar chart on the left shows that pneumonia, diabetes and hypertension have the highest number of hospitalizations.



5.0 Conclusions & Discussion

Our team developed a COVID risk and mortality prediction tool, leveraging a Random Forest model with 66% accuracy and a Tableau dashboard with Python APIs.While it's not without its limitations, its potential for refinement and expansion presents exciting avenues for future research and development.

5.1 Relevance & Impact

20% of COVID hospitalized patients need ICU care [18] thus the model is vital for healthcare workers, administrators, policymakers, and patients, providing quick risk and mortality predictions based on patient information. This aids in efficient treatment, risk stratification, and resource allocation during potential COVID spikes, benefiting patients with personalized care and improving outcomes. Public health officials and researchers can leverage the model's data to understand disease dynamics and intervention effectiveness across diverse populations.

5.2 Risks and Limitations

Our model is limited to Mexican patient data, making it non-universal. Success may lead to global expansion or separate models for each country. It is also crucial to acknowledge that machine learning models aren't infallible, and incorrect predictions may occur. Currently, our model employs binary classification, which may oversimplify the complexities of patient outcomes. However, the model aims to enhance decision-making for healthcare workers, improving treatment efficiency and quality.

5.3 Future Work

Moving forward, our model could be refined to address inherent limitations. Currently operating under a binary classification framework, adopting a probabilistic approach would offer more nuanced insights. This approach would provide a spectrum of values indicating the likelihood of outcomes, rather than strictly classifying them as 0 or 1. Such refinement would enable more detailed risk assessments and predictions, particularly valuable in healthcare. While binary classifiers are effective for clear-cut decisions, embracing a probabilistic approach could enrich our model's capabilities in healthcare decision-making.

Note: All team members will work collectively on each stage and have contributed equal efforts.

.

6.0 References

[1] "COVID Dataset," www.kaggle.com. https://www.kaggle.com/datasets/meirnizri/covid19-dataset

[2] M. Cascella, M. Rajnik, A. Aleem, S. C. Dulebohn, and R. Di Napoli, "Features, Evaluation, and Treatment of Coronavirus (COVID-19)," PubMed, 2024. doi: link

[3] D. A. Berlin, R. M. Gulick, and F. J. Martinez, "Severe Covid-19," New England Journal of Medicine, vol. 383, May 2020, doi: https://doi.org/10.1056/nejmcp2009575

[4] C. Bonand, "The Effect of Age on Mortality in Patients With COVID-19: A Meta-Analysis With 611,583 Subjects," Journal of the American Medical Directors Association, vol. 21, no. 7, pp. 915–918, Jul. 2020 doi: 10.1016/j.jamda.2020.05.045

[5] K. Romero Starke, D. Reissig, G. Petereit-Haack, S. Schmauder, A. Nienhaus, and A. Seidler, "The isolated effect of age on the risk of COVID-19 severe outcomes: a systematic review with meta-analysis," BMJ Global Health, vol. 6, no. 12, p. e006434, Dec. 2021, doi: https://doi.org/10.1136/bmjgh-2021-006434

[6] C. Hu et al., "Early prediction of mortality risk among patients with severe COVID-19, using machine learning," International Journal of Epidemiology, vol. 49, no. 6, pp. 1918–1929, Sep. 2020, doi: https://doi.org/10.1093/ije/dyaa171

[7] L. Kompaniyets, "Underlying Medical Conditions and Severe Illness Among 540,667 Adults Hospitalized With COVID-19, March 2020–March 2021," Preventing Chronic Disease, vol. 18, 2021, doi: https://doi.org/10.5888/pcd18.210123

[8] M. Treskova-Schwarzbach et al., "Pre-existing health conditions and severe COVID-19 outcomes: an umbrella review approach and meta-analysis of global evidence," BMC Medicine, vol. 19, no. 1, Aug. 2021, doi: https://doi.org/10.1186/s12916-021-02058-6

[9] C. I. Bloom et al., "Risk of adverse outcomes in patients with underlying respiratory conditions admitted to hospital with COVID-19: a national, multicentre prospective cohort study using the ISARIC WHO Clinical Characterisation Protocol UK," The Lancet Respiratory Medicine, Mar. 2021, doi: https://doi.org/10.1016/s2213-2600(21)00013-8.

[10] M. A. Quiroz-Juárez, A. Torres-Gómez, I. Hoyo-Ulloa, R. de J. León-Montiel, and A. B. U'Ren, "Identification of high-risk COVID-19 patients using machine learning," PLOS ONE, vol. 16, no. 9, p. e0257234, Sep. 2021, doi: https://doi.org/10.1371/journal.pone.0257234

[11] N. Casillas, A. Ramón, Ana María Torres, P. Blasco, and J. Mateo, "Predictive Model for Mortality in Severe COVID-19 Patients across the Six Pandemic Waves," Viruses, vol. 15, no. 11, pp. 2184–2184, Oct. 2023, doi: https://doi.org/10.3390/v15112184

[12] K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad, and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting COVID-19 mortality," BMC Medical Informatics and Decision Making, vol. 22, no. 1, Jan. 2022, doi: https://doi.org/10.1186/s12911-021-01742-0

[13] B. Chen et al., "A model to predict the risk of mortality in severely ill COVID-19 patients," Computational and Structural Biotechnology Journal, vol. 19, pp. 1694–1700, 2021, doi: https://doi.org/10.1016/j.csbj.2021.03.012

[14] Y.-D. Gao et al., "Risk factors for severe and critically ill COVID-19 patients: A review," Allergy, vol. 76, no. 2, pp. 428–455, Feb. 2021, doi: https://doi.org/10.1111/all.14657

[15] O. Albitar, R. Ballouze, J. Ping Ooi, and S. Maisharah Sheikh Ghadzi, "Risk Factors for Mortality among COVID-19 Patients," Diabetes Research and Clinical Practice, vol. 166, p. 108293, Jul. 2020, doi: https://doi.org/10.1016/j.diabres.2020.108293

[16] A. Vaid et al., "Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach," JMIR Medical Informatics, vol. 9, no. 1, p. e24207, Jan. 2021, doi: https://doi.org/10.2196/24207

.

[17] B. Musheyev et al., “Functional status of mechanically ventilated COVID-19 survivors at ICU and hospital discharge,” Journal of Intensive Care, vol. 9, no. 1, Mar. 2021, doi: https://doi.org/10.1186/s40560-021-00542-y

[18] S. S. Zakariaee, N. Naderi, M. Ebrahimi, and H. Kazemi-Arpanahi, “Comparing machine learning algorithms to predict COVID-19 mortality using a dataset including chest computed tomography severity score data,” Scientific Reports, vol. 13, no. 1, p. 11343, Jul. 2023, doi: https://doi.org/10.1038/s41598-023-38133-6

.