Longitudinal Prediction of the Need for Invasive Mechanical Ventilation in Hospitalized Patients with COVID-19 Infection Using Predictive Modeling

Introduction

The COVID-19 pandemic continues to be a major cause of disease and death across the globe, resulting in tremendous strain on hospital resources. Patients with COVID-19 infection can rapidly develop respiratory distress, and over one-third of patients requiring intubation and ventilator therapy progress to multi-organ failure and death (Lim et al., 2021). The SARS-CoV-2 pandemic has also been a logistical nightmare in resource management, and it has become imperative to effectively allocate ICU beds, staff and ventilators. Thus, there is an urgent need for the health care team at any hospital to accurately predict which patients are at risk for requiring Invasive Mechanical Ventilation (IMV), so that active measures can be taken to ameliorate negative outcomes.

Clinical experience and past studies have shown that the timing of treatment has a profound effect on disease outcomes by mitigating the cytokine storm and preventing the development of acute respiratory distress syndrome (Gupta et al., 2021). Previous studies have also demonstrated an increase in leukocyte and neutrophil counts to be associated with severe SARS-CoV-2 infection. ACE2 receptors on the pulmonary epithelium serve as entry receptors for the SARS CoV-2 virus. Pulmonary infiltration of neutrophils, with subsequent degranulation and production of interleukin-2 and cytokines, results in widespread organ damage and death. Once the cytokine storm ensues, destruction of pulmonary alveolar integrity spirals into failure of adequate oxygenation and results in the need for mechanical ventilation (Gibson et al., 2020).

This project aims to develop a neural network and Random Forest algorithm to predict the need for progression to IMV in patients by analyzing routinely collected laboratory data and comorbidities upon initial presentation to the emergency room at Stony Brook Hospital, New York to support clinical decision-making. The primary goal is to provide a way to identify high risk patients for early administration of dexamethasone and antiviral drugs such as remdesivir, which may ameliorate disease progression and reduce morbidity and mortality. Expensive therapies that specifically target SARS-CoV-2, such as convalescent plasma, sotrovimab, and monoclonal antibodies can be directed to patients with poor disease trajectories to mitigate an aggressive course and prevent long-term complications.

Methods

Study Population and Data Acquisition:

This project utilized the COVID-19 Persons Under Investigation registry of the Stony Brook Hospital ED from 7 February 2020 to 8 January 2022 to create two observational cohorts. There were 30,082 adult COVID-19 positive patients as determined by real-time polymerase chain reaction for coronavirus 2 (SARS-CoV-2), of which 18,288 were hospitalized. Patients who did not have full codes or were provided comfort care only were excluded. The final sample size included 193 patients who were hospitalized and placed on Invasive Mechanical Ventilation and 18,095 patients who were hospitalized and not placed on IMV. Structured Query Language (SQL) was used to collect demographic information, chronic comorbidities, laboratory tests, and vital signs. **Demographic data** included age, gender, ethnicity, and race. Comorbidities included obesity, diabetes, respiratory diseases, chronic kidney disease (CKD), heart failure, and coronary artery disease (CAD). **Vital signs** included systolic and diastolic blood pressure, respiratory rate, pulse oxygen saturation, and temperature. **Lab values** included creatinine, procalcitonin, aspartate aminotransferase (AST), alanine aminotransferase (ALT), ferritin, lactate dehydrogenase (LDH), white blood cell count, platelet count, C-reactive protein (CRP), lymphocyte count, D-dimer protein, brain natriuretic protein (BNP), and albumin. These clinical variables were collected upon initial presentation to the hospital.

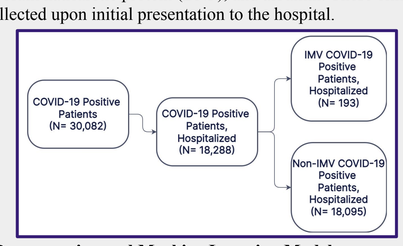

Fig. 1 - Flowchart detailing patient data in Montefiore database

Data Preprocessing and Machine Learning Models:

Seven suitable clinical variables were chosen to be used in machine learning analysis: diabetes, obesity, procalcitonin, AST, ALT, d-dimer, and ferritin. The target variable was invasive mechanical ventilation. SQL and Python (Pandas package) were used to compile the data into a table consisting of 7962 patients, with patients not requiring IMV serving as a control. Data was preprocessed to remove unrealistic outliers found in the lab value data, categorical data was transformed to numerical data (using the one-hot encoding machine learning method), and missing lab values were replaced with the median value. Lab values in the data set were filtered based on thresholds. An oversampling technique called SMOTE (an imblearn Python library) was used to artificially generate more IMV patients.

Neural networks are series of algorithms that mimic the operations of the human brain to recognize relationships between data. Each neural network consists of layers containing neurons, and neurons receive input from other neurons, process that input, and produce an output. Open-source software, namely Jupyter Notebook (Python), Tensorflow, and Keras were utilized to develop a neural network consisting of two types of layers: fully connected dense layers (*N=3*) and dropout layers (*N=2*). As the most frequently used layer, *dense layers* are pertinent to data classification. Each neuron in a dense layer receives input from all neurons in the layer immediately preceding it. Meanwhile, *dropout layers* randomly eliminate communications transmitted by its neurons not directly related to the training data. Dropout is commonly used to prevent overfitting, a type of modeling error that makes the ML model only accurate with the training set and not so when provided new data.

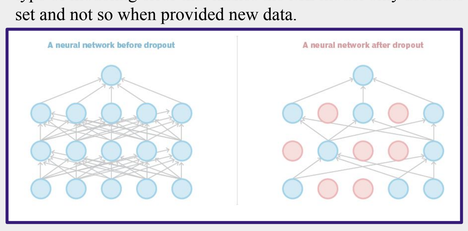

Fig. 2 - Model demonstrating the benefits of dropout layers by showing a neural network before and after dropout was used

The inputs for the neural network consisted of the clinical variables extracted for IMV patients and the control group (non-IMV patients). All data was extracted at the time of admission. The output (binary) was being placed on IMV, and the dataset was randomly split into 70% training data and 30% testing data. In other words, 70% of the data extracted from the database would be used to teach the model and improve model performance, while 30% would be used to test the model and analyze model performance.

The number of nodes, or neurons, in each layer of the neural network was specified. The input layer consisted of 7 nodes (the number of features in the training set), the first dense layer consisted of 128 nodes, the second dense layer had 32 nodes, and the third dense layer had 16 nodes, all using the rectified linear activation function (ReLU). ReLU is a piecewise linear function that will output the input directly if it is IMV positive. Otherwise, it will output zero.

Additionally, a 20% dropout rate was included for the first two dense layers. A sigmoid function in the output layer was used to decide whether or not a neuron should be activated, as this is a binary classification. Training was performed for 10 epochs with a batch size of 32. Epochs are hyperparameters that indicate the number of times that the learning algorithm will work through the entire training dataset. Meanwhile, the batch size refers to the number of training examples utilized in one training iteration. For this model, a learning rate of 0.001 and a classification threshold of 0.8 proved optimal. Both are tuning hyperparameters ('settings' of the network) that were adjusted until a desired output was reached. The model was compiled using the Adam optimizer, an efficient gradient descent algorithm. The clinical variables were ranked in order of importance using Sci-kit Learn and SHapley Additive exPlanations (SHAP), a Python package that explains the output and importance of ML models.

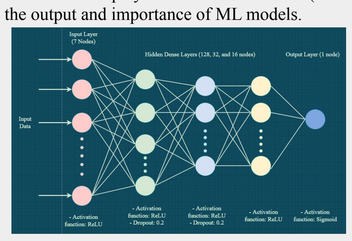

Fig. 3 - Model of constructed neural network (small dots represent many nodes)

$$R(z) = \begin{cases} z & z > 0 \\ 0 & z <= 0 \end{cases}$$

Fig 4 - Equation of ReLU linear function

$$S(z) = \frac{1}{1 + e^{-z}}$$

Fig 5 - Equation of sigmoid function

Results

Figure 7 summarizes the demographics, comorbidities, and average vital signs for hospitalized patients on IMV ($N = 193$) vs. hospitalized patients not on IMV ($N = 18{,}095$). Compared to the Non-IMV group, the IMV group had more males (*56.477%*) and African Americans (*37.373%*). As seen in other studies, the data indicated that patients in the IMV group were more likely to have comorbidities such as obesity, diabetes, respiratory disease, chronic kidney disease, heart failure, and coronary artery disease. All vital signs including systolic blood pressure, diastolic blood pressure, respiratory rate, pulse oxygen saturation, and temperature were not statistically different between groups in our data set, and thus were not utilized in the predictive models. Based on the metric of median age, hospitalized patients on IMV were 6.985 years older.

Figure 8 summarizes the average laboratory values for the IMV group vs the non-IMV group. AST, ALT, Ferritin, LDH, and BNP were statistically different between groups. Lymphocyte counts were virtually identical, with a difference of 0.1 between averages.

Characteristics	Invasive Mechanical Ventilation, Hospitalized N = 193	No Invasive Mechanical Ventilation, Hospitalized N = 18,095
Demographics, n		
Median Age (yrs.)	63.642	56.657
Male	109 (56.477%)	8155 (45.068%)
Female	84 (43.523%)	9940 (54.932%)
Hispanic	59 (30.570%)	7609 (42.050%)
Non-Hispanic Black	72 (37.373%)	6053 (33.451%)
Non-Hispanic White	32 (16.580%)	2282 (12.611%)
Non-Hispanic Other	15 (7.772%)	8552 (47.262%)
Comorbidities, n		
Obesity	120 (62.176%)	9573 (52.904%)
Diabetes	60 (31.088%)	4448 (24.581%)
Respiratory Disease	60 (31.088%)	2901 (16.032)
Chronic Kidney Disease	41 (21.244%)	2458 (13.584%)
Heart Failure	64 (33.161%)	2609 (14.418%)
Coronary Artery Disease	12 (6.218%)	373 (2.061%)
Vital Signs, avg		
Systolic Blood Pressure	125.775	126.742
Diastolic Blood Pressure	68.222	69.991
Respiratory Rate	21.525	20.284
Pulse Oxygen Saturation	97.000	97.428
Temperature	98.290	98.266

Fig. 7 - Table showing data regarding demographics, comorbidities, and vital signs for the IMV vs. no-IMV group

Characteristics	Invasive Mechanical Ventilation, Hospitalized N = 193	No Invasive Mechanical Ventilation, Hospitalized N = 18,095
Lab Values, avg		
Creatinine	2.046	1.933
Procalcitonin	3.932	2.616
Aspartate Aminotransferase	153.702	57.056
Alanine Aminotransferase	102.805	47.663
Ferritin	1958.659	1339.830
Lactate Dehydrogenase	572.051	439.749
White Blood Cell Count	10.830	9.048
Platelet Count	227.612	244.064
C-Reactive Protein	11.317	8.564
Lymphocyte Count	1.637	1.647
D-dimer Protein	7.558	4.441
Brain Natriuretic Peptide	3844.101	1900.863
Albumin	3.104	3.540

Fig. 8 - Table showing data regarding average lab values for the IMV vs. no-IMV group

Both the Neural Network Random and the Forest Model ranked Procalcitonin, ALT, AST, D-Dimer, Ferritin, diabetes, and obesity (in order of importance) as the top predictors of Invasive Mechanical Ventilation. Importance refers to how much the model takes each factor into consideration. The factor most strongly associated with IMV was Procalcitonin; this result was consistent across machine learning methods.

	features	importance
0	Procalcitonin_Value	0.614709
1	Alanine_Aminotransferase_Value	0.069970
2	Aspartate_Aminotransferase_Value	0.085058
3	D-Dimer_Value	0.105817
4	Ferritin_Value	0.098860
5	Diabetes_Yes	0.011964
6	Obesity_Yes	0.013622

Fig. 9 - Feature importance (how much weight each feature has in making a decision) for the neural network

	features	importance
0	Procalcitonin_Value	0.607344
1	Alanine_Aminotransferase_Value	0.074026
2	Aspartate_Aminotransferase_Value	0.083083
3	D-Dimer_Value	0.106042
4	Ferritin_Value	0.102296
5	Diabetes_Yes	0.011858
6	Obesity_Yes	0.015350

Fig. 10 - Feature importance (how much weight each feature has in making a decision) for the Random Forest algorithm

A confusion matrix was plotted for both models, where the number of correct and incorrect predictions is summarized with count values and broken down by each class. The classes included were:

- True Positives (bottom right) = The cases in which the model predicted YES and the actual output was YES.
- True Negatives (top left) = The cases in which the model predicted NO and the actual output was NO.
- False Positives (top right) = The cases in which the model predicted YES and the actual output was NO.
- False Negatives (bottom left) = The cases in which the model predicted NO and the actual output was YES.

The learning rate and classification threshold were adjusted to obtain optimal values.

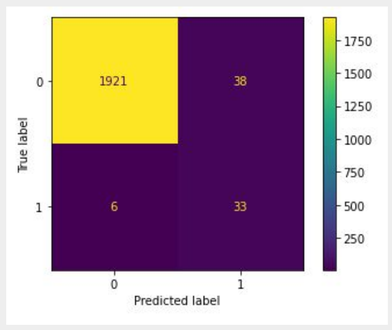

Fig. 11- Confusion matrix (constructed with Sci-kit Learn) for neural network

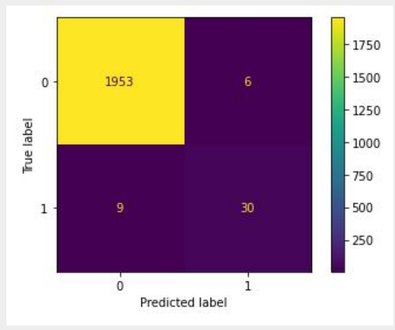

Fig. 12- Confusion matrix (constructed with Sci-kit Learn) for Random Forest algorithm

A classification report was developed for both models. The metrics used to classify each model are described below (True Positive/Negative and False Positive/Negative definitions explained in Confusion Matrix description).

- Recall = the measure of the model correctly identifying True Positives
- Precision = the ratio between the True Positives and all the Positives
- F-score = the harmonic mean between precision & recall
- Support = the number of occurrences of the given class in a dataset

Precision and recall are proper metrics to analyze a model trained from an imbalanced dataset, or a dataset where one category has a much higher count than the other.

	precision	recall	f1-score	support
0	1.00	0.98	0.99	1959
1	0.41	0.85	0.55	39
accuracy			0.97	1998
macro avg	0.70	0.91	0.77	1998
weighted avg	0.99	0.97	0.98	1998

Fig. 13 - Classification report (constructed with Sci-kit Learn) showing values regarding precision, recall, F1-score, and support for neural network

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1959
1	0.83	0.77	0.80	39
accuracy			0.99	1998
macro avg	0.91	0.88	0.90	1998
weighted avg	0.99	0.99	0.99	1998

Fig. 14 - Classification report (constructed with Sci-kit Learn) showing values regarding precision, recall, F1-score, and support for Random Forest algorithm

The Neural Network Model yielded an AUC (Area Under ROC Curve) of 0.88475, and the Random Forest Model yielded an AUC of 0.90624. The AUC provides an aggregate measure of performance across all possible classification thresholds (ranging from 0-1). A model whose predictions are never correct will have an AUC of 0.

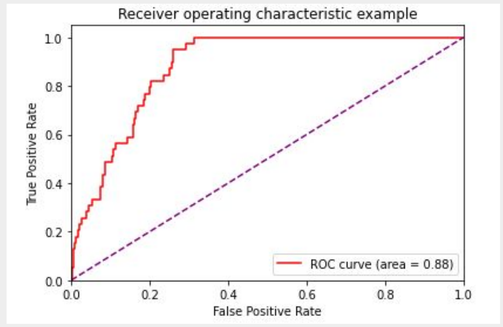

Fig. 15 - ROC curve graph (AUC at bottom right) for neural network

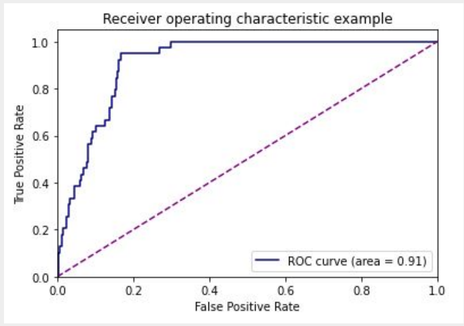

Fig. 16 - ROC curve graph (AUC at bottom right) for Random Forest algorithm

References

Burdick, H., Lam, C., Mataraso, S., Siefkas, A., Braden, G., Dellinger, R. P., McCoy, A., Vincent, J. L., Green-Saxena, A., Barnes, G., Hoffman, J., Calvert, J., Pellegrini, E., & Das, R. (2020). Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. *Computers in biology and medicine*, *124*, 103949. https://doi.org/10.1016/j.compbiomed.2020.103949

Lim, Z. J., Subramaniam, A., Ponnapa Reddy, M., Blecher, G., Kadam, U., Afroz, A., Billah, B., Ashwin, S., Kubicki, M., Bilotta, F., Curtis, J. R., & Rubulotta, F. (2021). Case Fatality Rates for Patients with COVID-19 Requiring Invasive Mechanical Ventilation. A Meta-analysis. *American journal of respiratory and critical care medicine*, *203*(1), 54–66. https://doi.org/10.1164/rccm.202006-2405OC

Garcia-Gordillo, J. A., Camiro-Zúñiga, A., Aguilar-Soto, M., Cuenca, D., Cadena-Fernández, A., Khouri, L. S., Rayek, J. N., Mercado, M., & ARMII Study Group (2021). COVID-IRS: A novel predictive score for risk of invasive mechanical ventilation in patients with COVID-19. *PloS one*, *16*(4), e0248357. https://doi.org/10.1371/journal. Pone.0248357

Li, W., Lin, F., Dai, M., Chen, L., Han, D., Cui, Y., & Pan, P. (2020). Early predictors for mechanical ventilation in COVID-19 patients. Therapeutic Advances in Respiratory Disease. https://doi.org/10.1177/1753466620963017

Lu, J. Q., Musheyev, B., Peng, Q., & Duong, T. Q. (2021). Neural network analysis of clinical variables predicts escalated care in COVID-19 patients: a retrospective study. *PeerJ*, *9*, e11205. https://doi.org/10.7717/peerj.11205

Nicholson, Christopher J et al. "Estimating risk of mechanical ventilation and in-hospital mortality among adult COVID-19 patients admitted to Mass General Brigham: The VICE and DICE scores." *EClinicalMedicine* vol. 33 (2021): 100765. doi:10.1016/j.eclinm.2021.100765

Gupta, A., Gonzalez-Rojas, Y., Juarez, E., Crespo Casal, M., Moya, J., Falci, D. R., Sarkis, E., Solis, J., Zheng, H., Scott, N., Cathcart, A. L., Hebner, C. M., Sager, J., Mogalian, E., Tipple, C., Peppercorn, A., Alexander, E., Pang, P. S., Free, A., & Shapiro, A. E. (2021). Early treatment for covid-19 with SARS-COV-2 neutralizing antibody sotrovimab. *New England Journal of Medicine*, *385*(21), 1941–1950. https://doi.org/10.1056/nejmoa2107934

"Diagrams.net - Free Flowchart Maker and Diagrams Online." *Flowchart Maker & Online Diagram Software*, https://app.diagrams.net/. Education, Ibm Cloud. "Random Forest." *IMB RF*, 26 Jan. 2021, www.ibm.com/cloud/learn/random-forest. Bushaev, Vitaly. "How Do We 'Train' Neural Networks ? - Towards Data Science." *Medium*, 22 Oct. 2018, towardsdatascience.com/how-do-we-train-neural-networks- edd985562b73.

Methods (Continued)

A machine learning Random Forest was employed using Sci-kit Learn. Along with a neural network, an RF model appeared to be the most suitable for the classification task. A random forest is a machine learning technique used to solve regression and classification problems. These algorithms consist of numerous decision trees, which are decision support techniques that form tree-like structures (see Fig). Simply put, Random Forest takes a random subset of the features, builds multiple decision trees, and merges them together to get a more accurate and stable prediction. This model utilized the ensemble learning method, which is a popular research hotspot in machine learning and is developed based on the single machine learning method. The ensemble learning method combines many classifiers to provide solutions to complex problems - it ensures the reliability of predictions along with the robustness of the ML model.

Bagging, or Bootstrap Aggregation, is the ensemble technique used by Random Forest and the one developed for this project. "Bagging" algorithms aim to reduce the complexity of models that overfit the training data. They take several weak models and aggregate the predictions to select the best prediction. The weak models specialize in distinct sections of the feature space, which enables bagging leverage predictions to come from every model to reach the final product.

The dataset was split into two subsets, the train and the test set. The train set was used for training the machine learning model, while the test set was used for the evaluation of its performance. During the training phase, the machine learning algorithm was fed with the train set (X_train and y_train). After the number of decision trees specified were constructed, test predictions were compared (y_test_pred) with the true values (y_test), which evaluated how well the model performed. The ensemble ML model was chosen because it efficiently tackled overfitting and performed better with noise data. The number of trees in the forest, or the n_estimators, had to be designated, and a larger number of decision trees typically correlates with higher accuracy. Thus, a count of five hundred was used.

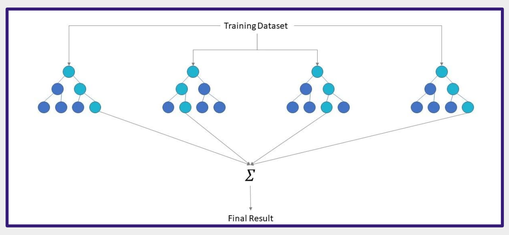

Fig. 6 - Diagram of a simple Random Forest with four decision trees

Performance Evaluation and Statistical Analysis:

Neural network data were split 70% for training and 30% for testing. Prediction performance was evaluated for both the neural network and the Random Forest Model by area under the curve (AUC) of the receiver operating characteristic (ROC) curve for the test data set. The average ROC curve and AUC were obtained (using Sci-kit Learn) with ten runs and standard deviations were obtained.

An accuracy score (to determine the number of samples correctly classified) was determined for both machine learning models, however, accuracy was not suitable to use because models trained from imbalanced datasets often produce extremely high accuracies (>99%). Thus, a classification report (y_test, y_test_pred) was taken for both models. It included several metrics for making comparisons, ranging from O-1. They included precision, recall, F1- score, and support. Confusion matrices consisting of four categories (True Positive/Negative, False Positive/Negative) were plotted for both models.

Discussion

Past studies have linked certain clinical data and demographic profiles to a worse prognosis in patients with COVID-19 infection. This data set (from Stony Brook Hospital) showed demographics, comorbidities, and lab values such as older age, male sex, diabetes, pre-existing lung disease, obesity, elevation in procalcitonin levels, C-reactive protein, and D-dimer to be predictors for invasive mechanical ventilation.

It was striking that this data set showed that vital signs on initial hospital presentation had minimal effect on the likelihood of IMV and were excluded from the machine learning models. A study from Massachusetts General Brigham Hospital (Nicholson, et al., 2021) described vital signs such as dyspnea, systolic blood pressure, as well as X-ray abnormalities, ALT, AST, direct bilirubin, ESR, ferritin, fibrinogen, LDH, MCH, and elevated white blood cell count to be risk factors for invasive mechanical ventilation. The cause for this variation in results is unclear but may be because of sample size or the demographic population in which the data was collected (The Bronx, NY).

Both machine learning models (NN and Random Forest Model) ranked the seven features used by degree of importance similarly, which suggests that these results are accurate. The factor most strongly associated with IMV was elevated Procalcitonin levels, with the two factors least strongly associated with IMV being diabetes and obesity. Another study from Huazhong University, China, (Wen Li, et al., 2020) found diabetes and obesity to significantly increase the need for IMV. However, both of the machine learning models found these two comorbidities to have an importance of < 0.2. A possible cause for this variation is that unlike the Chinese study, our data set was collected over a two year period ending in January 2022, which will capture the changes in clinical practice patterns. Age was not a significant predictor of ventilation need, which dispels the commonly held belief that COVID-19 only causes severe disease in the elderly. Current data suggests that people under 40 are common vectors of COVID-19. While there is clear evidence that young patients are less likely to die from COVID-19, once hospitalized, they commonly require IMV for prolonged time periods, which is a huge drain on hospital intensive care resources (Burdich et al., 2020).

The neural network and the Random Forest model performed similarly, with a difference in AUCs of 0.0215 and identical precision and support scores. The differences in the weighted averages of recall and F1-scores for the two models were slight enough to be considered negligible. Confusion matrices were also nearly identical, with the neural network having a slightly large number of False Positives. Again, accuracy did not serve as a suitable metric for measuring performance as the dataset was imbalanced. Both machine learning models performed well and can serve as suitable algorithms to predict a patient's likelihood for IMV. The neural network model, however, likely has more potential to be improved in the future and is more advanced. Tuning hyperparameters such as learning rates, classification thresholds, the batch size, and the number of epochs and layers. These can be adjusted accordingly in the future to achieve more optimal results.

Possible errors that may have arisen regarding the models likely involved tuning hyperparameters or the data set being limited to a specific geographic region. Small tweaks were consistently made to both models until the most suitable results were achieved, but (as the chosen values were not perfect) further adjustments and additions could still be made to the neural network. For the Random Forest algorithm, the maximum depth of the decision trees, number of n_estimators, and the maximum number of features to be included at each node split in the tree could also be changed to improve performance. In data preprocessing, thresholds were placed on lab values and null values were filled in, likely worsening the performance of both models. Constructing algorithms with new, updated and more complete data would increase model validity, however, new samples were unable to be added to the dataset. In the future, different types of ML models such as Ensemble boost algorithms, feature engineering (adding new features), and other oversampling techniques (artificially generating new data) could be explored.

Conclusion

The goal of this project is to assist physicians in effectively triaging high risk patients who present to the emergency room. Results demonstrate that both predictive models *can* be used in the emergency department of hospitals as a simple way to augment clinical decision making. Early administration of antivirals and steroids and closer monitoring of patients with a high likelihood of needing IMV/ICU care may result in preventing deaths and saving valuable resources. This risk stratification will potentially allow for these patients to be treated via more aggressive but potentially expensive modalities and conserve limited resources by not utilizing them unnecessarily. The weighted importance of clinical variables that are indicative of a patient's chance for IMV may also prove to be crucial in clinical decision making.

A significant strength of this project is a much larger sample size relative to previous studies (N=7962). Another strength is that the data reflects evolving changes in the need for IMV based on current changes in clinical practice.

Future work will need to be done to improve the accuracy of the predictive models. Another possibility is to expand the use of predictive modeling to outpatient settings to help guide clinic staff on which patients need to be directed to the ER for more aggressive therapy. Additional studies are also needed to determine the effect of individual treatment modalities on the likelihood of needing ICU care or IMV. The machine learning models are also limited to the New York area - the findings need to be replicated in a geographically diverse multi-institutional setting.