Forecasting Hospital Bed Capacity

ZhiLei Xiao

501012295

Derya Kici

April 4, 2022

Table of Contents

[Abstract 3](#_Toc98107617)

[Significance 4](#_Toc98107619)

[The Literature 4](#_Toc98107620)

[Methodology and Data Descripion 8](#_Toc98107620)

Exploratory Data Analysis [10](#_Toc98107620)

Autoregressive Integrated Moving Average Modelling and Prediction [15](#_Toc98107620)

Feed Forward Neural Network Modelling and Prediction [19](#_Toc98107620)

Long Short-Term Memory Modelling and Prediction [21](#_Toc98107620)

Conclusion [23](#_Toc98107620)

[References 26](#_Toc98107622)

# Abstract

Public Health Ontario is an online government resource which provides province-wide health data for study. The aim in making the information available is to allow individuals and organizations to conduct scientific analyses which can in turn be used to inform policies and practices to improve citizen’s health.

One such area of interest is the ability of hospitals to provide adequate care to those who need it. While hospitals require a certain amount of hospital beds to be kept available for patient use, the Ontario Ministry of Health mandates the number of beds offline at any given time (Ontario, 2020). This is especially true after situations such as overfilled hospitals, following the third wave of COVID-19. There is a certain amount of hospital beds that are kept online and thus, available, in a unit at a time, each requiring their own allocation of resources to maintain the status of preparedness for a patient to occupy at a moment’s notice. Hospitals may also have a number of beds offline to ensure efficiency of resource usage so that the cost of maintaining unused available beds does not impede on the function of the unit. What this study seeks to find is the optimal number of beds to keep online in Ontario Intensive Care Units (ICU) during the COVID-19 pandemic. The aim is to achieve this by using known methods of machine learning on historical time series data from the ICU occupancy dataset (Ontario, 2020) gathered by the Ministry of Health Ontario 2020 to predict the days when high volumes of patients requiring critical care emerge. Univariate time series analysis and forecasting using the ARIMA model, and the neural networks inspired by the work of Zhang (2003) will be used to answer the research question of “What are the daily occupancy demands of Ontario's critical care units during the covid pandemic?”.

### Objectives

The objectives of this project are:

* To explore and analyze hospital ICU occupancy during the pandemic in Ontario
* To identify patterns and trends in ICU hospitalizations along the progression of the virus
* To develop machine learning models that predict the daily occupancy demands of Ontario's critical care units using patterns found in earlier stages of the pandemic
* Provide recommendations for hospitals to increase their preparedness for high patient volumes and needs in the future

### Significance

If occupancy patterns of patients for these hospital beds can be tracked, studied, and understood, predictions can be made more accurately for the most efficient number of beds ready to go online at any time. This will be particularly useful when these predictions can be made in advance to anticipate rush times or downtimes to make best use of hospital resources. In this way, machine learning can be useful. Specific data of past historical occupancy rates for patients with COVID, and data during the different stages of the pandemic, can be used in time series data to try and accurately predict the amount available of beds needed overall. In addition, results of this machine learning can help set a stronger precedent for preparations in cases of future epidemics, natural disasters, and extreme cases of hospital overflow.

# The Literature

#### Industry Background

The type of literature referenced is firstly from a field or industry perspective, looking to provide insight on the infrastructure of hospitals and how they approach capacity management, as well as Covid-19 news and reports which can be used to provide more insights on the applications of the research question. From the background provided by the 2016 report of care in Canadian hospitals, the overall state of intensive hospital care has been increasing in necessity due to the aging overall population and an increase in illness severity. While this report may be used to draw a general background understanding it was conducted in 2016 during regular non pandemic times. I expect the ICU data (Ontario 2020) during the COVID-19 period would have an upward shift in volume across the board as well as more dramatic increases of ICU occupancy during surges, behaviors such as plateaus, and decreases as vaccines roll out to the population.

Interestingly the Canadian Institute for Health information's report on December 9th, 2021, regarding COVID-19 impact on hospital services reports add that with each wave two and wave three hospital inpatient admissions were actually lower in comparison to the pre-pandemic period. However, the report also states that non COVID-related admissions for procedures were delayed and cancelled with the rising numbers of COVID infections. Respiratory conditions were found to rise in accordance with infection rate as well as ICU admission and ventilator demand increased.

More recent statistics in the Weekly Epidemiological Summary for COVID-19 in Ontario: Focus on January 30, 2022, to February 5, 2022, indicates an overall decrease in outbreaks. Public Health Ontario does warn, however, that the ICU admission and death rates may be underreported due to detection that may not have happened in time before death or infection after follow-up completion. Public Health Ontario’s January 31, 2022, report *on Omicron in Ontario: Risk Analysis for Approaching Public Health Measures in Winter 202*2 predicts a peak in Omicron infection around 4 weeks, states that currently infection rates are low but warns that there is massive uncertainty regarding the trajectory of the variant in Ontario specifically highlighting higher risk level during the winter of 2022. It is important to examine the wider context of the pandemic in relation to the ICU hospital occupancy in that there is the implication that the trends found in the anticipated overall state and infection rate of the virus would affect ICU rates; during infection outbreaks the volume of patients is expected to rise and along with it, a corresponding increase in ICU beds to meet the demand will be necessary.

#### Literature Applications of Time Series Methods in Hospitals

Secondly, academic literature focusing on the applications of predictive data analytics on time series with relevant methodology in the medical field is consulted. The research question of finding the daily occupancy demands of Ontario's critical care units specifically attempts to address patterns in ICU capacity during the pandemic. In some sense, it will help consolidate what hospitals already know and struggle with (in terms of ensuring the correct number of beds available) into relevant data that can be used to make quantitative predictions. It will give insight not only on the rate of patients going in and out of intensive care, but also overall patients’ needs as the coronavirus evolves.

Comparing this proposed methodology with what other practitioners have done in the consulted literature for the purpose of this endeavor, the research pertaining to the specific aim of forecasting ICU bed capacity during a pandemic is unavailable at this time. However, several researchers have conducted related studies, Jones et al. (2009) used the Vector Autoregression (VAR) model to forecast the demands and to analyze the relationship between 8 different resources within the emergency department at three different hospitals. While VAR was shown to be effective in Jones et al.’s study, it would not be as suitable of an approach for this specific research due to the nature of the research question and dataset required. Jones et al. (2009) required datasets which had multiple variables for different hospital resources which could be used to study their relationships among the variables as they change over time unlike the dataset for this project which only focuses on one variable changing upon itself throughout time. However, it does illuminate the option and provides insight on the path of using univariate VAR as a forecasting method which could be applied to this research question.  Furthermore, the study by Batal et al. (2001) uses stepwise linear regression to isolate "significant" variables, using which the authors developed their own prediction equation to forecast patient visits depending on weather conditions. In addition to the multivariate differences in Jones et al. (2009), the methods in Batal et al. (2001) also cannot be applied in the same way, and because a main focus of the topic is to use machine learning, the specifically developed prediction equation would not apply either. There is no doubt, however, that provided with a different dataset with more recorded details on outside variables such as weather, hospital resources, and/or patient information in addition to ICU occupancy it would be beneficial and worth analyzing through employing the methods used in Jones et al. (2009) and Batal et al. (2001) papers as they have shown how to include and consider the influences of multiple variables in their forecasting methodology.

Two studies examining predictive analytics in hospital care units report similar results in their model usage. Kim et al. (2014), and Tandberg and Qualls (1994), both used ARIMA methods for forecasting.

The two main papers used to guide the project are Zhang (2003) and Khashei and Bijari (2010). In these studies, the authors both examine the use of the autoregressive integrated moving average (ARIMA) model and the artificial neural networks (ANN) for time series forecasting. The study by Zhang (2003) illuminates the strengths in which the ARIMA model has been diploid as well as the strength and circumstances in which ANN have been used and found to be most effective and appropriate. They then combine the two methods in an attempt to utilize the advantages in each model to compensate for the other model’s shortcomings. His research and insight on the models’ state that ARIMA is linear in its prediction of future values and are typically constrained to linear behaviors in past observations. Due to the nonlinearity of some datasets, the performance of linear prediction models such as ARIMA may not most effectively capture those nonlinear behaviours in its forecasting. To address the issue of nonlinearity, Zhang (2003) suggests the use of ANN. they assert that artificial neural networks R one type of model suitable for “approximating various nonlinearities in the data.” (Zhang, 2003).

## Methodology and Data Description

The working dataset is the [*Availability of Adult and Pediatric ICU Beds and Occupancy For Covid-Related Critical Illness (CRCI)*](https://data.ontario.ca/dataset/availability-of-adult-icu-beds-and-occupancy-for-covid-related-critical-illness-crci) obtained from the Ontario Data Catalogue. It is a government site that provides datasets for the whole province of Ontario. The dataset contains observations which were recorded on a daily basis for the period of May 01, 2020, to February 14, 2022.

The variables include the date and eight variables measuring the number of persons, adult or pediatric in the following four categories:

* number of patients in ICU/pediatric ICU for COVID-related critical illness (CRCI)
* number of patients in ICU/pediatric ICU for non-CRCI reasons
* number of patients ICU/pediatric ICU beds that are unoccupied
* total number of patients in ICU/pediatric ICU for any reason

The overall methodology consists of determining the topic of finding the demand for critical care ICU beds for adult patients with CRCI during the COVID pandemic in Ontario, then moving on to narrowing down the dataset, cleaning the dataset of duplicates or missing values, selecting the feature variable, conducting exploratory data analysis which includes examining the properties of the data, testing for stationarity and differencing if non-stationary, then dividing in to training and test sets for the modelling component. Finally during the modelling component the data will be modelled and forecast three times with the models, autoregressive integrated moving average (ARIMA), the feed forward neural network, and the long short-term memory (LSTM) model.

Diagram

Description automatically generated

Figure 1 Overall methodology

A brief examination of the descriptive statistics of the dataset in the attached Github link finds the following Figure 2 which shows the values of the first and five days of the data indicating the period and the starting and ending values. Table

Description automatically generated

Figure 2 First five observations of initial four dataset variables.

Table

Description automatically generated

Figure 3. Last five observations of main variable of interest

**Exploratory Data Analysis**Table

Description automatically generated

Figure 4. Descriptive statistics table for dataset variables

While there are four variables to the dataset, the main category of focus is the number of non-pediatric patients in the ICU for CRCI. This is because in terms of attribute elimination, we will want to focus on the patients in critical conditions rather than amount of beds available or total because it will be more useful to know now many are absolutely necessary to set aside for CRCI patients within the ICU, and the focus will be on adult CRCI patients as the adult population including seniors has been most fatally affected by COVID. Therefore, the main variable of focus will be adult CRCI patient numbers. As the objective is univariate time series analysis, the statistical method means all other attributes can be eliminated. This means the dataset will be subsetted into a univariate time series with the date as the index and the number of non-pediatric patients in the ICU for CRCI as the variable. A brief examination of the total number of patients in ICU for any reason is conducted to contextualize the data in the real world in order to check for province-wide shortages. This is because these best represent the quantity of those who require high resources ventilators in CRCI, and those who still need to be in the ICU regardless of infection status. Figure 2 shows the basic descriptive statistics drawing attention to the first column which will read the amount of adults in CRCI were recorded 655 across 655 days, mean of 251.78 people, median of 172 people,  and standard deviation of 204.27 shows that dispersion of the data relative to its mean is not necessarily very spread out nor very clustered around the mean and the distribution is further explored using boxplot, histogram, and interquartile ranges below.

The variable is of numeric integer type. The data is quantitative and discrete as it measures each patient occupancy in terms of beds and there are no half-patient or a fraction of a bed type of occupancy. The finding that there were no zeros in the number of available ICU or pediatric beds at any time during the period, implying that overall, across Ontario there the occupancy of ICUs were not exceeded. Despite this, it has been repeatedly reported that hospitals were at maximum occupancy. A possible explanation for this could be that hospitals in more populated locations in Ontario were experiencing shortages while less populated areas had an excess. Including the hospital names or locations in the dataset could have helped combat this constraint and would be more helpful in finding ways to allocate resources throughout the province.

  Moving forward in the data cleaning process the check for missing values is conducted and found zero missing values. The check for duplicate observations also concluded that there were no duplicates.

Chart, line chart

Description automatically generated

Figure 5. Line graph depicting adult CRCI patients in the ICU

A box plot and histogram is created to analyze the distribution of the observations. The histogram is divided in terms of months to analyze the distribution of data points per month which found that the most and highest points were in May and April with also the largest spread and lowest during August. The interquartile range is 235 with the upper whiskers at 702 and lower at -238, however since there is no such thing as negative occupancy the lower whiskers end at 0. This shows that there is variability in comparison with the full range of the data being between 18 and 889 and while some of the data is distributed between 114.5 and 349.5, much of it lies beyond in the upper ranges. There are 35 outliers residing in the upper range and the upper whiskers extend quite high. This reflects the spike in patients experienced during the months of May and April in 2021.

Chart, box and whisker chart

Description automatically generated

Figure 6. Boxplot showing distribution of data points per month.

Chart, box and whisker chart

Description automatically generated

Figure 7. Box and whisker plot showing the overall distribution of data points for adult CRCI patients in the ICU.

Time series data can be broken down into different components separating its level, trend, seasonality, and residual. These components of the time series that have consistency or reoccur in the series could be shown and modeled. The level describes the average value, the trend describes the propensity of the value to increase or decrease over time. Seasonality is the repeating short-term cycles found in the time series, and the residual or noise describe the remaining contributors towards the time series which cannot be explained or modeled by the previous components. When these components are added together in order to create the time series it is said that the time series is an additive model whereas when the components are multiplied it is multiplicative. Results show that trend exists in both models, not much seasonality in both, considerable noise in the additive model meaning more variability than in the multiplicative model. These models are expressed and shown in the decomposition of the dataset as follows:

Additive: Time Series  = Level + Trend + Seasonality + Residual(also known as Noise)

Multiplicative: Time Series  = Level \* Trend \* Seasonality \* Residual(also known as Noise)

Chart, histogram

Description automatically generated

Figure 8. Additive decomposition of the time series

Chart, histogram

Description automatically generated

Figure 9. Multiplicative decomposition of the time series

As a pre-emptive measure to the modelling of the time series a stationarity check will be conducted to ensure that the mean and the variance of the dataset is constant, thereby determining whether removing the differences of levels in the dataset or making it stationary is necessary. This is a necessary step for the implementation of the Auto-Regressive Integrated Moving Average (ARIMA) model used further on during the modelling portion of the study.

Checking for stationarity is conducted through the use of the Augmented Dickey Fuller test (ADF Test) in which the null hypothesis is that the data is not stationary with an alternative hypothesis that the data is stationary. ADF test returned a p-value of 0.137 which is larger than the p-value of 0.05 needed to be able to reject the null hypothesis therefore the data is not stationary.

## Autoregressive Integrated Moving Average Modelling and Prediction

The first part of the ARIMA model is the autoregressive (AR) portion in which the variable of interest is forecasted using a linear combination of the variable’s past values. Autoregression describes the regression of the variable against itself.



Figure 10. Autoregressive model of order p (Hyndman & Athanasopoulos, 2018).

The parameter p is the autoregressive property of the model describing the lagged term or how far back in the variable past values will be used for the autoregression.

The second portion of the ARIMA model is the moving average (MA) portion that uses the past data point’s error terms to predict a current and future point’s value.

A picture containing text, clock

Description automatically generated

Figure 11. Moving average model of order q (Hyndman & Athanasopoulos, 2018).

The order ***q*** in the model represents previous values which are used to predict the current observation.  can be thought of as a weighted moving average of the past few forecast errors

The integrated part of the ARIMA model represented by order ***d*** is the degree of differencing done on the dataset. Differencing refers to the removal of trend from the dataset and to the degree of which is necessary to achieve stationarity. Stationarity is required to prevent the trend negatively impacting the prediction accuracy of the model, since the trend impacts values of data points at different times, stationarity is needed to create the independence so that the model can predict using past data points independent of the expected values across time periods, which results in more accurate predictions. The change between each ensuing data point in differencing can be shown as:



Figure 12. Integration of order d (Hyndman & Athanasopoulos, 2018).

For the implementation of the ARIMA model, the training of the model will be done on the first portion of the dataset of 571 observations in order to predict on the last quarter of 3 months which will be the test set. This is calculated in terms of weeks so 743 = 84 days for the test split and 655- 84 = 571 for the train portion.

The ADF test is then conducted on the train set to check for stationarity on the specific portion of the time series to determine whether integration will be necessary. The test returned a p-value of 0.14 which is less than 0.05 hence there is non-stationarity and integration will be necessary.

Hyperparameter tuning using auto ARIMA is used to determine the optimal values of p, d, and q which were (1,1,1) in order to train the model which was found to be more effective than manually differencing then modelling using ARMA.

 Akaike Information Criterion (AIC) was used to evaluate model selection. AIC assesses how well the fit of the model is on the data by estimating the prediction error of the models as compared to the data points. Auto ARIMA finds the optimal parameters for the model by minimizing AIC.

Two methods of making the data stationary were tested, one by manually differencing before modelling, meaning that the actual ARIMA modelling done was actually ARMA since d = 1 was pre-processed. The ARMA model of (1,0,1) resulted in an AIC score of 3777.427.

Directly using auto ARIMA with train set where in the integration/differencing was done by the model itself being and ARIMA of order (1,1,1) resulted in an AIC score of 3775.453 The second method results had a slightly higher AIC score so directly using auto ARIMA on train was the better method by 1.974.

After plotting the model on the actuals of the training data to check how well it fits, plot predict is conducted using the model to predict on the training set.

Chart, line chart

Description automatically generated

Figure 13 ARIMA model forecast.

Evaluations metrics used to determine how well the model predicted the last quarter of the data compared to the actual points of the test set are root mean-squared error (RMSE) and mean absolute percentage error. RMSE measures the standard deviation of the errors or residuals thus it shows the distance of how spread out from the actual series’ data points are from the model’s predicted values. MAPE measures the accuracy of the model prediction as a percentage.

The ARIMA model produced an RMSE of 376.451 and MAPE of 100%, which indicates that the model did not predict very well when evaluated on the test set. While the ARIMA model had a much lower run time of two seconds, it is not the most accurate in its predictions as compared to the subsequent models explored. This could be due to the unpredictable spikes that came along as time progressed and is not in line with the literature. Although Zhang (2003) did write that ARIMA did not do as well in capturing the non-linearities of time series as one of the limitations and that is congruent with the findings in this case.

## Feed Forward Neural Network Modelling and Prediction

Feed forward neural networks are a type of neural network wherein there is no feedback between the layers. Feed forward neural networks are a type of supervised deep learning algorithm which utilize interconnected nodes called neurons that accept data inputs which are evaluated in layers to produce outputs to estimate the relationships present in the data. The nodes are arranged in layers where the inputs are placed in the input layer and move forward to the hidden layer of nodes where each neuron has an activation function that evaluates the information to determine whether there exists a strong enough relationship to fire to the next neuron. This forms associations can be activated in recognition of a relationship within the data and allows for the neural network to learn and make predictions in a similar mechanism to biological neurons in the brain.

Training the neural network is the method by which known data is fed into the neural network and ran through the layers onto a known output and the accuracy of the prediction on the actuals is evaluated. This is conducted throughout several iterations or epochs until the neural network is determined to be predicting outputs well.

The feed forward neural network in this study begins by converting the data points into floating numbers to keep accuracy as the next step is to normalize the values. This is to ensure that the training of the neural network is not negatively affected by the magnitude of the values as it better learns when the points do not have large differences in the numbers and can evaluate on portion. The data is split into the train and test split of the year’s quarter as kept in consistency with the ARIMA model. Arrays of x and y are created as per the “to\_sequences” function in order to generate the inputs which are the arrays of seven data points that consists of x to predict the next one point of y this is to train the network by having it take in an input of one week to then predict the next day.

The x and y components are then fed into the build of the model using the rectifier or Rectified Linear Unit (ReLU) activation function which was chosen for its fast and effective performance with nonlinearities and sparse data, as properties that the time series in question does exhibit (Glorot, et al., 2011). Fitting the model and training is done for 20 epochs and the mean squared error (MSE) measuring how close the predictions are to the actuals for each epoch is plotted and the earliest model around 1 is shown to drastically minimize the MSE and is chosen and used to predict on the test set. The results of the feed forward model are 18.390 RMSE and 4.456% MAPE with a run time of around 4 seconds which is quite good in comparison to the ARIMA model. This is also in line with the literature in Zhang (2003) and Khashei and Bijari (2010) as it was found that the feed forward neural network worked better with the nonlinearities in the data and predicted well on the actuals as expected since the hospital capacity data has a considerable amount of nonlinearity.

Chart, line chart

Description automatically generated

Figure 14. Feed forward neural network forecast

## Long Short-Term Memory Modelling and Prediction

Long Short-Term Memory (LSTM) is a subset of recurrent neural networks. It is a type of neural network is similar to the feed forward model explored previously, however LSTM differs in many ways. The main difference is that in feed forward neural networks the inputs only move from node to node in one direction while LSTM has feedback between its connections. Through this method LSTMs can keep previous information in order to contextualize current data in relation to its predecessors. There are loops within the structure of the LSTM layers that facilitate the operation of keeping relevant previous data while discarding unrelated information and this allows LSTMs to work well with sequential data as it remembers the order and relationships of past data points as it relates to the current one. LSTMs specifically combat the issue of the vanishing gradient problem in recurrent neural networks where the weight of the network diminishes during the repeated backpropagation when training recurrent neural network models and becomes too small to contribute to the learning of the model. This is done through the structure of LSTM cells which are smaller network units within the LSTM. Each of these units consist of a cell, an input gate, an output gate and a forget gate. The gates determine which data in a sequence is important to keep or discard and by doing so learns to use relevant information to make predictions. The cell stores values over time intervals and the three gates control the information flow in and out of the cell (Yu, et al., 2019).Diagram

Description automatically generated

Figure 15. Architecture of an LSTM cell with a forget gate by Yu, et al. (2019).

The steps conducted in the setup of the modelling and prediction of LSTM is very similar to that of the feed forward mode since LSTM is also a type of neural network. The majority of the differences is conducted within the hidden layers of the neural network itself’ inner workings.

The preprocessing of normalizing the values, split of train and test data by the last quarter, x and y arrays of input representing inputting seven days to predict on the next day is fed into the model using rectifier activation function. The LSTM prediction on the actuals of the test data was not as accurate as in the feed forward model with an RMSE of 324.701 and an MAPE of 61.301%. This may indicate that remembering the sequential nature of the data for this particular time series was not as integral in affecting the results of future data points as necessary to use an LSTM and that the back propagation of the LSTM may have negatively impacted the training by remembering things that needed to be forgotten. Figure 17 summarizes the evaluation metrics of each model in table format for easier comparison.

Chart, line chart

Description automatically generated

Figure 16. LSTM forecast.

## Conclusion

**Evaluation Table**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ARIMA** | **FF NN** | **LSTM** |
| **RMSE** | 376.451 | 18.390 | 324.701 |
| **MAPE** | 100.083 | 4.456 | 61.301 |
| **Running Time** | 2.072 seconds | 4 seconds | 41 seconds |

Figure 17. Evaluation Table.

In summary, in order to explore and analyze hospital ICU occupancy during the pandemic in Ontario with a focus on adults in the critical care unit suffering from covid-related critical illness, data analysis was employed in the form of exploratory data analysis (EDA) to examine the available properties of the observations and what they may indicate and then further machine learning techniques such as ARIMA, and deep learning models such as feed forward neural networks and LSTM was employed to identify patterns and trends in ICU hospitalizations for adults with CRCI along the progression of the virus and use those relationships to make predictions in anticipation of future demands.

EDA revealed that historically highest hospitalization rates occurred during the spring and winter months which could suggest an impact of the weather possibly contributing to serious Covid sickness and that the distribution of the data has a large spread skewing toward the upper range. There is a large spike in occupancy during 2021 May, otherwise there is a slight upward trend to the time series and no seasonality.

In the resulting information from EDA and in line with the literature where ARIMA and artificial neural networks were found to be most effective in predicting time series data with respect to occupancy rates in healthcare organizations, ARIMA and feed forward neural networks along with LSTM were used to predict the last quarter of adult CRCI occupancy in the ICU. The findings were mostly congruent in comparison to those found within the literature. The feed forward neural network performed the best, with LSTM and ARIMA having the most deviations with the predicted versus the actual values. While LSTMs were not conducted within the literature, ARIMA results were stated to perform better with less nonlinearity in the data and shown to do so in the literature which is also in line with the apparent nonlinearity of the adult CRCI time series data. The difference in performance between the LSTM and feed forward networks, while both outperforming ARIMA, can be interpreted as a possible over complication in the attempts to model the data as LSTM was taking into account and remembering the sequential relationships in the data points to avoid the deterioration of weights in the training and perhaps the forgetting or releasing of certain data or having a higher bias in which data progresses through the network is needed in the case of predicting this particular time series.

Shortcomings of the methodology include the limitations of data in its collection and the somewhat short time span of two years. The analysis on the data is done on a single variable, hence, the univariate time series analysis. This was chosen due to the lack of some more information or data collection and can be improved through collection of other possible crucial predictors such as location and more specific data on the hospitals could yield more insight as indicated in the portion of EDA where occupancy was found to never have been at its limit for the whole of Ontario, however the hospitals in areas with higher population density definitely reported shortages of space and multivariate time series analysis may help account for this to also improve the distribution of resources for future outbreaks. The time span of May 01, 2020, to February 14, 2022, is almost two years wherein there are hints of spikes during certain periods, perhaps if the time period was longer spanning several years, certain cycles can be found to more accurately predict the spikes and lulls which could also be matched up with vaccine and booster rollouts.

As it stands, through the results of the EDA and machine learning with ARIMA, feed forward neural networks, and LSTM, there is the general consensus within the findings that a downward trend is anticipated so recommendations for hospitals to increase their preparedness for high patient volumes in the ICU specifically for CRCI illness and needs in the future is not of the most crucial at the moment, however allocation of resources to areas in Ontario with higher population density may be considered during the spring and winter months.

**References**

Batal, H., Tench, J., McMillan, S., Adams, J., & Mehler, P. S. (2001). Predicting patient visits to an urgent care clinic using calendar variables. *Academic Emergency Medicine*, *8*(1), 48-53.

CIHI Portal. [2016]. Care in Canadian ICUs. Ottawa, ON: *Canadian Institute for Health Information*. Retrieved from <https://secure.cihi.ca/free_products/ICU_Report_EN.pdf>.

CIHI Portal. [2016]. COVID-19’s impact on hospital services. Ottawa, ON: Canadian Institute for Health Information. Retrieved from https://www.cihi.ca/en/covid-19-resources/impact-of-covid-19-on-canadas-health-care-systems/hospital-services

CP 24. [2022]. Ontario reports 21 new COVID-19 deaths; ICU occupancy now exceeds 400. Retrieved from <https://www.cp24.com/news/ontario-reports-21-new-covid-19-deaths-icu-occupancy-now-exceeds-400-1.5733238?cache=%3FautoPlay%3Dtrue%3FclipId%3D104062%3FautoPlay%3Dtrue>

CTV News. [2021]. Ontario ICU occupancy could hit 200 patients in January as COVID-19 cases rise, modelling suggests. Retrieved from <https://toronto.ctvnews.ca/ontario-icu-occupancy-could-hit-200-patients-in-january-as-covid-19-cases-rise-modelling-suggests-1.5663239>

Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315-323). JMLR Workshop and Conference Proceedings.

Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

Jneny. (2022). Hospital Capacity. *Github*. Retrieved from https://github.com/Jneny/Hospitalcapacity.git

Jones, S. S., Evans, R. S., Allen, T. L., Thomas, A., Haug, P. J., Welch, S. J., & Snow, G. L. (2009). A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of biomedical informatics*, *42*(1), 123-139.

Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with applications*, *37*(1), 479-489.

Kim, K., Lee, C., O’Leary, K., Rosenauer, S., & Mehrotra, S. (2014). Predicting patient volumes in hospital medicine: A comparative study of different time series forecasting methods. *Northwestern University, Illinois, USA, Scientific Report*.

Ontario Ministry of Health. [2020]. *Availability of adult and pediatric ICU beds and occupancy for COVID-related critical illness (CRCI)*. Retrieved from <https://data.ontario.ca/dataset/availability-of-adult-icu-beds-and-occupancy-for-covid-related-critical-illness-crci>

Public Health Ontario. [2022]. WEEKLY EPIDEMIOLOGICAL SUMMARY. Retrieved from <https://files.ontario.ca/moh-covid-19-weekly-epi-report-en-2022-02-05.pdf>

Science Table. [2022]. Ontario Dashboard Tracking Omicron. Retrieved from <https://covid19-sciencetable.ca/ontario-dashboard/#effectivereproduction>

Tandberg, D., & Qualls, C. (1994). Time series forecasts of emergency department patient volume, length of stay, and acuity. *Annals of emergency medicine*, *23*(2), 299-306.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270. Retrieved from https://asset-pdf.scinapse.io/prod/2944851425/2944851425.pdf

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175.