Predicting Mechanical Ventilation Duration Using Deep Learning Models: A Comparative Analysis*

1st Huang Zhiwen
Electrical & Computer Engineering
National University of Singapore
Singapore
e0376960@u.nus.edu

***Abstract*—In this study, the goal was to predict mechanical ventilation (MV) duration for patients using deep learning models. A dataset consisting of 36 patients, some of which had multiple intubations, was carefully preprocessed to handle missing data and outliers. Only three features were selected for prediction due to the presence of many missing values in other features: Ventilator Reading End-Tidal Carbon Dioxide (ETCO2), Ventilator Reading Respiratory Rate (breaths/min), and Ventilator Reading Tidal Volume (VT) Exhaled. Group Shuffle Split was employed to split the dataset into training and validation sets, and a 5-fold cross-validation was used to evaluate model performance. Two baseline models were used: the mean value of the training set and a decision tree regressor. Four deep learning models were compared against the baselines: Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Temporal Convolutional Networks (TCN), and Gated Recurrent Units (GRU). The CNN model demonstrated the best performance in predicting mechanical ventilation duration. The results suggest the potential of deep learning techniques, particularly convolutional neural networks, for predicting mechanical ventilation duration in critical care settings. Limitations of the study include a small dataset with missing data and some intubations being re-intubations, which may affect the predictions.**

***Index Terms*—Mechanical Ventilation Duration Prediction, Deep Learning, Convolutional Neural Networks, Long Short-Term Memory, Gated Recurrent Units, Temporal Convolutional Networks, Time Series Data, Data Preprocessing, Group Shuffle Split, Cross-Validation, Model Comparison, Critical Care**

I. Introduction

Mechanical ventilation (MV) is a critical intervention for patients experiencing respiratory failure or requiring support during surgery. Accurate prediction of MV duration is essential for optimizing patient care, resource allocation, and reducing the risk of complications associated with prolonged ventilation [1]. Despite the importance of this prediction task, existing approaches often rely on clinicians' expertise, which may be subjective and can lead to inconsistencies in patient management [2].

Recent advances in deep learning techniques have demonstrated remarkable success in various healthcare applications, including time series analysis and prediction tasks [3]. However, the literature specifically focusing on MV duration prediction using deep learning methods remains limited. In light of these developments, this study aims to investigate the potential of deep learning models in predicting MV duration based on a limited set of patient time series data.

Several studies have explored the application of machine learning to predict mechanical ventilation outcomes or duration. However, it is still unclear if machine learning contributes to a higher extubation success rate [4]. Figueroa-Casas et al. developed predictive models to determine if a mechanical ventilation is prolonged or not, but they did not predict the exact duration [5] [6]. Mohammed Sayed utilized machine learning to predict the duration of mechanical ventilation for ARDS patients, but the results were not good enough for practical applications and deep learning models were not used [1].

In this paper, I analyze the performance of several deep learning models, including Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Temporal Convolutional Networks (TCN), and Gated Recurrent Units (GRU), on a dataset of 26 patients requiring MV, with some having multiple intubations. The dataset originates from Ministry of Health Holdings (MOHH) and contains information on 36 patients, of which 26 are included in the analysis after filtering out patients with a lot of missing or incomplete data. Due to the presence of many missing values in other features, only three features were selected for prediction: Ventilator Reading End-Tidal Carbon Dioxide (ETCO2), Ventilator Reading Respiratory Rate (breaths/min), and Ventilator Reading Tidal Volume (VT) Exhaled.

I employed Group Shuffle to split the dataset into training and validation sets, ensuring that no patient's data was present in both sets simultaneously. A 5-fold cross-validation was used to evaluate model performance. As baseline models, I employed the mean value of the training set and a decision tree regressor. The performance of the LSTM, CNN, TCN, and GRU models was compared against the baselines.

The CNN model demonstrated the best performance in predicting MV duration, highlighting the potential of deep learning techniques, particularly convolutional neural networks, for predicting MV duration in critical care settings using limited feature sets. This study contributes to the growing body of research on deep learning applications in healthcare and provides valuable insights for future research and clinical practice.

II. Dataset

A. Data Source and Description

The dataset used in this study was collected by doctors from MOHH. It consists of data from 36 patients who underwent mechanical ventilation in intensive care units (ICUs). The dataset covers various time periods, with each patient having different durations in the ICU. Some patients have multiple intubations, while others do not have any intubation records. A histogram in Figure 1 was used to show the distribution of MV duration.

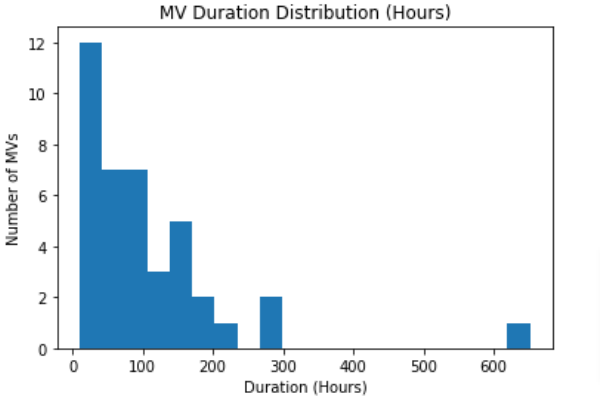


Fig. 1: MV Duration Histogram.

B. Data Preprocessing

During the preprocessing stage, several steps were taken to clean and organize the dataset. Firstly, I used a pandas dataframe to store the dataset. Duplicated rows were removed, as the dataset contained multiple duplicated data, especially for patients with multiple ICU admissions and surgeries. Rows with null values were also dropped. This cleaning process helped ensure the data quality for further analysis. Then the data was sorted based on 'Case Number', 'Created Datetime', and 'Item Description' columns. Since some patients had multiple intubations, each intubation was separated for model training purposes. Additionally, a new column, 'Hours', was added to represent the time of sampling the data for each intubation, ensuring that the time started from zero for each intubation. Out of the 36 cases, 9 cases (case 2, 3, 4, 12, 13, 18, 20, 33, 36) were removed from the dataset because they did not have ventilator data. Case 01 was also excluded due to its MV duration of over 600 hours, which was considered an outlier. Lastly, intubations with an MV duration of less than 10 hours were removed, as they were too short for training with the sliding window method.

C. Feature Selection

Although the ventilator machine gives many reading and setting data, only three features were selected for prediction. The reason for selecting these three features is that other ventilator reading and setting data have a lot of missing values, making them unsuitable for model training. The features selected as predictors are:

- **Ventilator Reading ETCO2**: End-tidal carbon dioxide (ETCO2) is the partial pressure or maximal concentration of carbon dioxide (CO2) at the end of an exhaled breath, which is considered to be a noninvasive estimate of the arterial carbon dioxide (PaCO2) level in the body [7].
- **Ventilator Reading Respiratory Rate (breaths/min)**: Respiratory rate (RR) is the number of breaths that a person takes per minute. It is usually measured when a person is at rest and simply involves counting the number of breaths for one minute by placing a hand on the chest or abdomen and counting each time the chest rises.
- **Ventilator Reading VT Exhaled**: Exhaled tidal volume (VT) is the amount of air that is exhaled during each breath. It is usually measured using a spirometer, which measures the amount of air that a person breathes in and out of their lungs [8].

These three predictors were also confirmed by doctors that they may be helpful for MV duration prediction as they are relevant indicators of lung function and the effectiveness of MV support.

III. Methodology

A. Train and Validation Data Split

In order to evaluate the performance of the deep learning models, I split the 40 MVs extracted from the dataset into training and validation sets. To maintain the integrity of the dataset and ensure that MVs belonging to the same patient do not appear in both the training and validation sets simultaneously, I employ a Group Shuffle Split approach. This method groups the MVs by their associated patient and shuffles them, maintaining the patient grouping throughout the process. I set the test size to be 0.3, which means 30% of the data is used for validation, while the remaining 70% is used for training the models. To further assess the effectiveness and robustness of the models, I also employ a 5-fold cross-validation technique. This means that the dataset is divided into five equally-sized folds, and the model is trained and evaluated five times, each time using a different fold as the validation set and the remaining four folds as the training set.

By using this approach, I minimize the risk of overfitting and can better generalize the model's performance across different patients. This allows me to assess the effectiveness of the models in predicting mechanical ventilation duration for patients in a real-world setting.

B. Sliding Window Method

To capture the temporal dependencies and patterns within the mechanical ventilation data, I employ a sliding window method to preprocess the data before feeding it into the deep learning models. The sliding window method enables me to analyze the variations in the data and uncover the inherent relationships between the features over time.

Given the sample rate of the data is one hour, I set the sliding window's slide to one hour as well. This means that the

window moves forward one hour at a time, extracting a time series from the data at each step. The sliding window method is applied separately to the training and validation sets after the train-validation split has been performed. This ensures that no information leakage occurs between the two sets during the preprocessing step.

In order to determine the optimal window size for the models, I experiment with different window sizes and analyze the influence of the window size on model performance. By selecting the appropriate window size, I can ensure that the models effectively capture the temporal dependencies in the data and provide accurate predictions for mechanical ventilation duration. The experiments results visualized as plots and shown in IV.

C. Baseline Model

In this study, I employed two baseline models to evaluate the performance of deep learning models in predicting the duration of mechanical ventilation. These baseline models provide a point of comparison to ensure that the deep learning models are effectively learning from the data and identifying relevant patterns.

1) Mean Value of the Training Set: The first baseline model is the mean value of the mechanical ventilation duration in the training set. This model represents a simple and intuitive guess of the MV duration. By comparing the performance of the deep learning models to this baseline, it can be determined whether the models are genuinely learning from the data and discovering patterns that contribute to better predictions.

2) Decision Tree Regressor: The second baseline model is the Decision Tree Regressor. A Decision Tree Regressor is a non-parametric supervised learning method that can be used for regression tasks. It works by recursively splitting the input space into non-overlapping regions and fitting a constant model (e.g., the mean) within each region. The tree structure provides a flexible and interpretable representation of the relationships between input features and the target variable [9].

The Decision Tree Regressor can be used as a baseline regression model for time series data because it can capture complex, non-linear relationships between input features and the target variable without making any assumptions about the underlying data distribution. Although it may not be as powerful as more sophisticated deep learning models in capturing temporal dependencies, it provides a reasonable point of comparison to assess whether the deep learning models are indeed leveraging the time series nature of the data to make more accurate predictions.

D. Deep Learning Models

In this section, four deep learning models are introduced: LSTM, CNN, TCN, and GRU. For all models, the Adam optimizer is used for optimization and the mean absolute error (MAE) serves as the loss function. Early stopping and learning rate reduction are employed as callbacks during training to ensure optimal model performance. The early stopping callback restores the best model weights when the validation loss stops improving, while the learning rate reduction callback decreases the learning rate when the validation loss plateaus. The model structures presented here are the result of experimentation with different architectures, starting from a single layer and incrementally adding layers and units to fine-tune each model for the best performance. The model structures are presented in Figure 2,3,4,5 respectively.

1) LSTM Model: The Long Short-Term Memory (LSTM) model is a type of recurrent neural network (RNN) that is capable of capturing long-range dependencies in time series data [10]. In this work, an LSTM model with 64 hidden units is used, followed by a dropout layer with a rate of 0.5 to prevent overfitting. The model then feeds into a dense layer with 32 units and a ReLU activation function, and finally, another dense layer with a single output unit and a linear activation function.

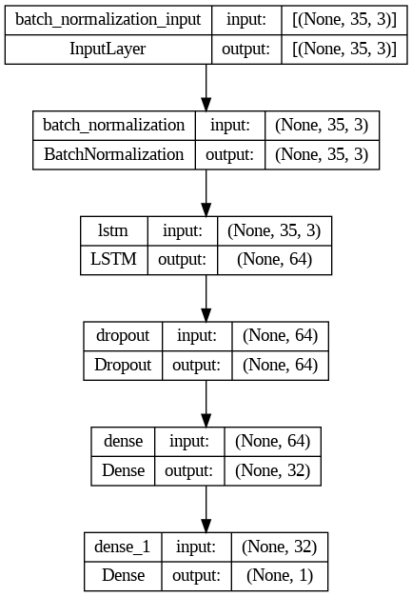


Fig. 2: LSTM Model Structure (With a sliding window size equal to 35).

2) CNN Model: The Convolutional Neural Network (CNN) model is used to capture local patterns in time series data [10]. The CNN model consists of a 1D convolutional layer with 64 filters and a kernel size of 3, followed by a max-pooling layer with a pool size of 2, and a dropout layer with a rate of 0.3. The model then includes a flatten layer, a dense layer with 32 units and a ReLU activation function, another dropout layer with a rate of 0.3, and finally a dense layer with a single output unit and a linear activation function.

3) TCN Model: The Temporal Convolutional Network (TCN) model is designed to capture both local and global patterns in time series data [11]. The TCN model uses dilated causal convolutions to handle long-range dependencies in the data. In this work, a TCN model with 32 filters, a kernel size of 3, and a dilation rate that increases exponentially from 1 to 256 is used. The model also includes skip connections to

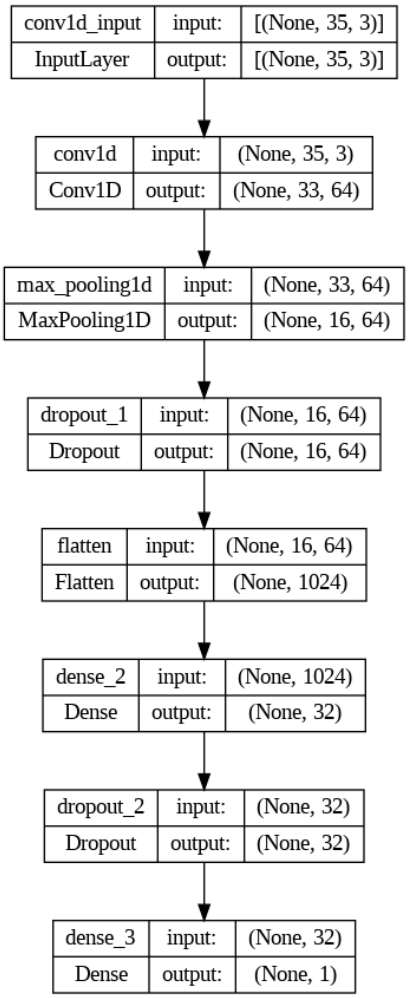


Fig. 3: CNN Model Structure (With a sliding window size equal to 35).

improve training efficiency. The model then feeds into a dense layer with a single output unit and a linear activation function.

4) GRU Model: The Gated Recurrent Unit (GRU) model is another type of RNN that is capable of capturing long-range dependencies in time series data [12]. The GRU model is used with 32 hidden units, followed by a dense layer with a single output unit and a linear activation function.

IV. Results and Analysis

In this section, the performance comparison of the six models (two baselines and four deep learning models) for ten different window sizes: 15, 20, 25, 30, 35, 40, 45, 50, 55, and 60. For each window size, box plots were created to visualize the distribution of the MAE across the five cross-validation folds. These box plots allow for an assessment of the variability in performance and the stability of the models for each window size. Considering the page limit, only the boxplot of models' MAE when window size = 55 and 60 are shown in Figure 5.

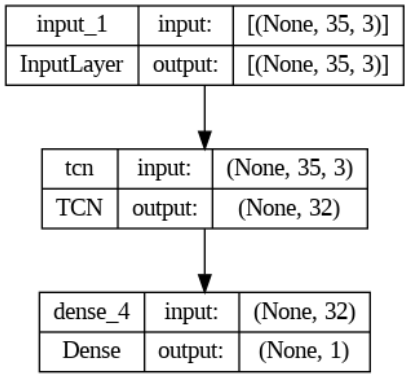


Fig. 4: TCN Model Structure (With a sliding window size equal to 35).

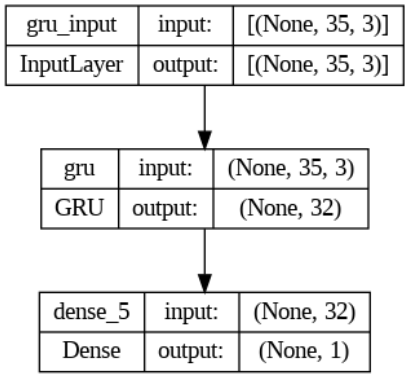


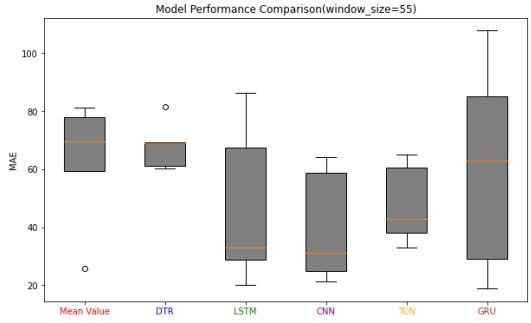
Fig. 5: GRU Model Structure (With a sliding window size equal to 35).

Additionally, the mean MAE of all models of different window sizes is shown in Table I. The best result of each window size was highlighted in the table. To visualize these results, it was plotted in Figure 6. This plot demonstrates the impact of window size on the performance of the models and allows for a comparison of their performance across different window sizes.

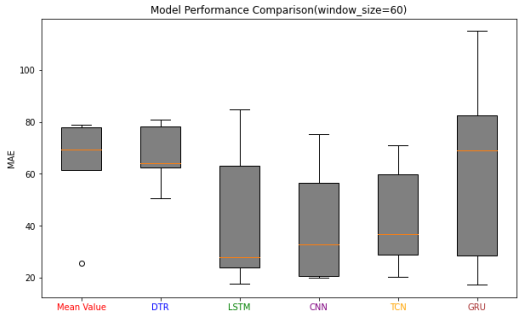
TABLE I: Mean MAE for each model and window size

Window Size	Mean	DTR	LSTM	GRU	CNN	TCN
15	64.62	77.17	**49.41**	54.57	50.75	54.57
20	64.46	77.06	48.84	53.70	**48.13**	52.39
25	64.43	73.95	53.71	53.87	**49.05**	51.15
30	64.26	70.55	51.17	53.23	**46.05**	48.27
35	63.88	71.01	**45.09**	53.60	46.87	48.80
40	63.49	74.37	**44.45**	54.38	45.40	51.20
45	63.08	72.09	48.53	54.88	**44.33**	51.55
50	62.85	66.14	**41.28**	56.13	44.80	47.33
55	62.81	66.83	43.99	61.77	**42.98**	48.86
60	62.59	67.81	45.42	62.90	**37.06**	49.03

In the results, it can be observed that all deep learning models outperform the two baseline models. Among the deep learning models, CNN achieves the best overall performance, as demonstrated by its small mean MAE across different window sizes. Especially, when the window size is increased



(a) window size = 55



(b) window size = 60

Fig. 6: Boxplots of the average mean absolute error (MAE) of six models (two baselines and four deep learning models) across five cross-validation folds for window sizes equal to 55 and 60.

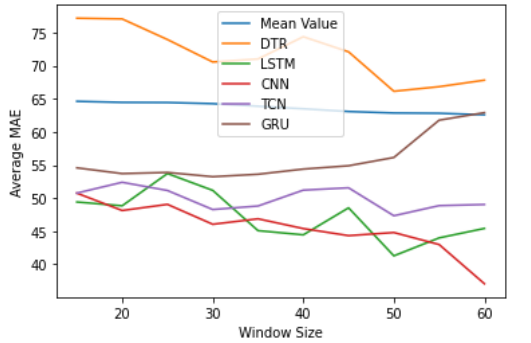


Fig. 7: The impact of window size on the average mean absolute error (MAE) of each model. This line plot illustrates the relationship between the window size (ranging from 15 to 60) and the average MAE across five cross-validation folds for the six models (two baselines and four deep learning models).

to 60, the mean MAE of CNN is 37.06, which is 41% less than the MAE of the mean value baseline. TCN also exhibits strong performance, with a particularly noteworthy aspect being the focused MAE distribution across different folds. This implies that the standard deviation of TCN's performance is small, making it a more stable model.

The LSTM model performs between CNN and TCN, it performs worse than TCN when window size is small but outperforms TCN when window size is larger than 35. On the other hand, GRU model, although generally better than the baseline models, displays unstable performance across folds, resulting in a large standard deviation. Interestingly, GRU's performance does not improve with increasing window size; in fact, it appears to worsen. This is in contrast to the other three deep learning models, which demonstrate better performance with larger window sizes.

This observation suggests that larger windows, which contain more information, enable the models to better identify patterns and dependencies within the data. However, this trend does not hold for the GRU model, indicating that there may be an underlying issue with its ability to utilize the additional information provided by larger window sizes effectively. Overall, the analysis highlights the strengths and weaknesses of each model and their performance in relation to window size, offering valuable insights for further model optimization and selection. To visualize the predictied results of the CNN model, a figure was plotted when the window size was set to 60, as shown in Figure 8. The figure shows the actual MV duration and predicted MV duration with the sliding window as the input. It is worth noting that there is no patient overlap between the train and validation data, meaning the model has never seen the actual values before. From the figure, it can be observed that the CNN model has learned the internal pattern and has a very similar trend to the actual data, indicating the good performance of the model. Interestingly, upon closer

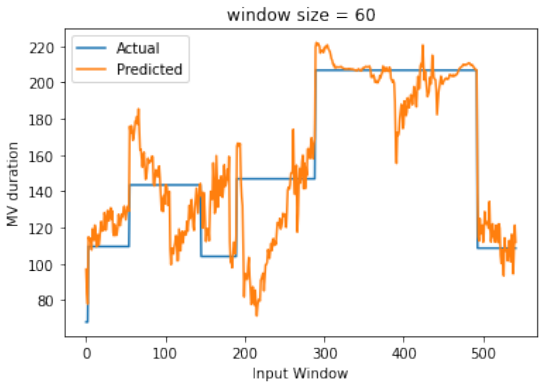


Fig. 8: Comparison of Predicted and Actual Values for the CNN Model

examination, it was observed that the predicted values for each MV duration in the figure were decreasing. This may indicate that the model predicted the total MV duration calculated from the input window. This hypothesis could be verified by further investigation in future work.

V. Conclusion and Limitations

In conclusion, this study investigated the use of various deep learning models, including CNN, TCN, LSTM, and GRU, to predict the duration of mechanical ventilation for patients. The results demonstrated that all deep learning models outperformed the two baseline models, with CNN generally showing the best performance. It was also observed that LSTM and

TCN models performed well, potentially due to their ability to capture temporal dependencies and patterns within the time series data. These findings suggest that deep learning models can be effectively employed to predict mechanical ventilation duration, offering valuable insights for clinicians and healthcare providers.

However, several limitations should be considered when interpreting the results of the study. Firstly, the dataset is relatively small, with only 26 patients available for analysis. A limited sample size can reduce the generalizability of the findings and may lead to overfitting in the models, potentially affecting the performance of the models when applied to larger, more diverse patient populations.

Secondly, the study only utilizes three features from the ventilator machines, as a substantial amount of reading and setting data is either missing or not recorded. The limited availability of these features may hinder the models' ability to fully capture the underlying patterns and dependencies within the data. The performance of the models could potentially be improved if additional relevant features were included in the analysis.

Lastly, the dataset consists of both successful and unsuccessful mechanical ventilation events, with some patients requiring re-intubation. This variability in the success of intubations may affect the models' ability to accurately predict the duration of mechanical ventilation. Future studies could potentially benefit from distinguishing between successful and unsuccessful intubations or focusing solely on successful cases to better understand the factors that contribute to the accurate prediction of mechanical ventilation duration.

Despite these limitations, the study provides a valuable starting point for future research aiming to improve the prediction of mechanical ventilation duration using deep learning techniques. Further studies could explore the incorporation of additional features, larger datasets, and more diverse patient populations to enhance the generalizability and performance of the models. Additionally, future research could investigate the impact of differentiating between successful and unsuccessful intubations on the accuracy of the predictions. Researchers could further explore the prediction of the remaining MV duration from the input time point rather than the total MV duration. This approach may offer a more accurate representation of the medical pattern within the ventilator data.

REFERENCES

[1] Sayed, Mohammed, David Riaño, and Jesús Villar. "Predicting Duration of Mechanical Ventilation in Acute Respiratory Distress Syndrome Using Supervised Machine Learning." *Journal of Clinical Medicine* 10.17 (2021): 3824. https://doi.org/10.3390/jcm10173824.

[2] Taylor, Fiona. "A comparative study examining the decision-making processes of medical and nursing staff in weaning patients from mechanical ventilation." *Intensive and Critical Care Nursing* 22.5 (2006): 253–263. https://doi.org/10.1016/j.iccn.2005.11.001.

[3] Esteva, Andre, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Kuan Chou, Cindy Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. "A guide to deep learning in healthcare." *Nature Medicine* 25.1 (2019): 24–29. https://doi.org/10.1038/s41591-018-0316-z.

[4] Igarashi, Yutaka, Keita Ogawa, Kohei Nishimura, Shin Osawa, Hitoshi Ohwada, and Soutaro Yokobori. "Machine learning for predicting successful extubation in patients receiving mechanical ventilation." *Frontiers in Medicine* 9 (2022): 252. https://doi.org/10.3389/fmed.2022.778889.

[5] Figueroa-Casas, Juan B, Sean M Connery, Ricardo Montoya, Alok K Dwivedi, and Sang Lee. "Accuracy of Early Prediction of Duration of Mechanical Ventilation by Intensivists." *Annals of the American Thoracic Society* 11.2 (2014): 182–185. https://doi.org/10.1513/AnnalsATS.201306-195OC.

[6] Figueroa-Casas, Juan B, Alok K Dwivedi, Sean M Connery, Reginald Quansah, Lisa Ellerbrook, and Jannet Galvis. "Predictive models of prolonged mechanical ventilation yield moderate accuracy." *Journal of Critical Care* 30.3 (2015): 502–505. https://doi.org/10.1016/j.jcrc.2015.01.020.

[7] Hunziker, Sabina, Michelle Bivens, Michael N Cocchi, and Anne C Miller. "End-tidal and arterial carbon dioxide measurements correlate across all levels of physiologic dead space." *Respiratory care* 55.3 (2010): 288–293. https://doi.org/10.1016/j.iccn.2005.11.001.

[8] Hallett, Sasha, Fadi Toro, and John V. Ashurst. "Physiology, Tidal Volume - StatPearls - NCBI Bookshelf." *StatPearls Publishing*, 2022. https://www.ncbi.nlm.nih.gov/books/NBK482502/.

[9] Loh, Wei-Yin. "Fifty years of classification and regression trees." *International Statistical Review* 82.3 (2014): 329–348. https://doi.org/10.1111/insr.12016.

[10] Taye, Mesfin Mulugeta. "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions." *Computation* 11.3 (2023): 52. https://doi.org/10.3390/computation11030052.

[11] Nan, Mingxi, Mihai Trăscău, Ana-Maria Florea, and Codruta C Iacob. "Comparison between Recurrent Networks and Temporal Convolutional Networks Approaches for Skeleton-Based Action Recognition." *Sensors* 21.6 (2021): 2051. https://doi.org/10.3390/s21062051.

[12] Gharehbaghi, Ahmad, Reza Ghasemlounia, Farzaneh Ahmadi, and Mohsen Albaji. "Groundwater level prediction with meteorologically sensitive Gated Recurrent Unit (GRU) neural networks." *Journal of Hydrology* 612 (2022): 128262. https://doi.org/10.1016/j.jhydrol.2022.128262.