Economy based Multivariate Oil price prediction using LSTM

ABSTRACT

Predicting unrefined petroleum prices is a crucial exercise for decision-makers, financiers, and academics inside the energy industry. In this review paper, we present an economic system-based multivariate oil price forecasting model that utilizes LSTM, a sophisticated deep learning algorithm that could capture tricky and nonlinear interactions in data with temporal order. Our model takes WTI crude price and numerous economic variables, containing US Dollar Index Futures, Gold Futures, Ten-year US Bond Yield, and S&P 500, as input characteristics. Before model training, we executed data preprocessing to eliminate outliers and enhance model accuracy. via exploratory data analysis, we diagnosed and indifferent outliers between 2007 and 2009, which were influenced by financial crisis. We also expelled data from 2020 and after, as the hurricane of COVID-19 should distort our judgments. Our findings display that the LSTM version has powerful predicting talents and can correctly forecast oil price bases on economic indicators. This review paper also stresses the significance of macroeconomic indicators in oil price forecasting.

***Index Terms*—** LSTM, crude oil, WTI

1. INTRODUCTION

Petrol, diesel fuel, and polymers are just a few of the items made from crude oil, a valuable commodity. A multitude of variables, such as supply and demand, prevailing economic circumstances, and geopolitical developments, affect the price of crude oil. Crude oil price forecasting is a difficult task [1] that calls for comprehensive models that can capture the intricate and nonlinear interactions between these variables. Using a machine learning model is one way to forecast the price of crude oil. In order to discover the links between the numerous elements that affect the price of crude oil, machine learning models are trained using historical data. Once trained, a model may be used to forecast crude oil prices in the future.

The LSTM is a promising [2] model for time series data in finance. It differs from a typical recurrent neural network in that it has a unique topology where memory cells in lieu of regular nodes in the buried layers. Because of its distinctive structure, LSTM has stronger fitting capabilities. Each of the cells that make up an LSTM has an intricate internal structure. Three different types of gates, an internal state, and a number of inputs make up a cell's internal structure as seen in Fig1. Weighted values for the connection signals are represented by the connections between the gates and nodes. These weighted values are crucial for artificial neural networks since choosing weights is the most crucial factor in LSTM. For predicting oil prices using a multivariate time series analysis, LSTM is employed here because it has a significant processing capability for data with temporal order.

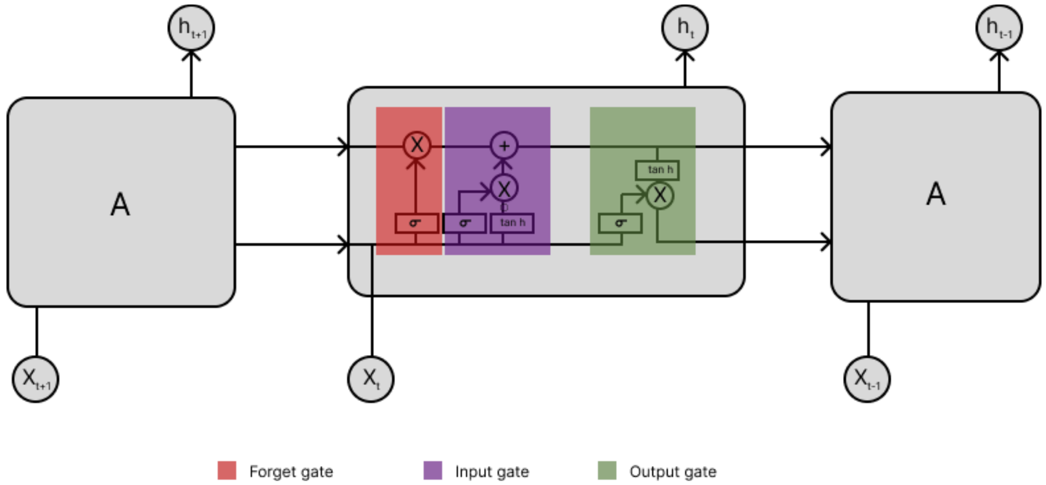

Fig-1 LSTM structure [3]

2. LITERATURE SURVEY

A new method for forecasting crude oil prices using artificial neural networks is presented in the paper by Nalini Gupta and Shobhit Nigam [1]. By determining the ideal lag and quantity of delay effects that regulate crude oil prices, the authors emphasise the key benefit of their strategy, which is the capacity of the ANN to continually capture the unstable pattern of crude oil prices.

To get the most precise and close findings, the authors also touch on the significance of adjusting the latency over a period of time. The findings obtained using the suggested model greatly exceed those obtained using other approaches, with a best Root Mean Squared Error (RMSE) of 7.68.

In order to forecast the price of West Texas Intermediate (WTI) crude oil, the study "Evolutionary Neural Network model for West Texas Intermediate crude oil price prediction" suggests an alternate strategy based on a genetic algorithm and neural network (GA-NN). **[6]**The authors show that their GA-NN technique surpasses baseline approaches in terms of prediction accuracy and computing efficiency by comparing it to them.

The WTI crude oil price predicted by the suggested GA-NN model and the observed price are both statistically equivalent, according to the authors' Mann-Whitney test, which they also used to compare. Potential uses for the suggested model include the creation of worldwide crude oil price estimation, development strategies, and industrial production policies.

The study of a cutting-edge method that combines a genetic algorithm with a neural network makes a significant addition to the field of crude oil price prediction. The benefit of this method is that it uses genetic information to adjust the neural network parameters, increasing prediction accuracy. The R2 score of 0.91722, a strong measure of the model's capacity to capture the variance in crude oil prices, is another way in which the authors illustrate the viability of their strategy.

The performance of the Multi-recurrent Network (MRN) in predicting crude oil prices over a range of prediction horizons is assessed in the study by O. Orojo, et. al [7]The MRN is a recurrent neural network model that demonstrates intricate, flexible, and rigid state-based memory. In order to explicitly represent the shocks in oil prices brought on by the financial crisis, the authors compare the MRN with various models, including Feedforward Multi-layered Perceptron (FFMLP), Simple Recurrent Network (SRN), and Long Short-Term Memory (LSTM).

To assess the models, the authors combine out-sample data from October 2003 to March 2015 with in-sample data made up of important indicator variables collected across the pre-financial crisis era (July 1969 to September 2003). They discover that the MRN performs better than the FFMLP, SRN, and LSTM models in predicting crude oil prices, especially when it comes to simulating shocks brought on by the financial crisis. Five years before the 2008 financial crisis, the MRN was able to identify significant latent characteristics buried in the input signal. This implies that the indicator variables may be used as early warning signs of future financial disturbances.

The research makes a significant addition to the field of crude oil price forecasting by proving the efficacy of the MRN, a very straightforward yet potent recurrent neural network model. Policymakers and market participants who want to comprehend the factors influencing crude oil prices and make wise decisions based on their estimates should consider the authors' conclusions.

In order to increase the accuracy of crude oil price predictions, further research has expanded on the work described in this study by looking at additional MRN method adjustments or switching to completely other machine learning approaches. Using the MRN methodology suggested by the authors of this research as an example, a recent study titled "Deep learning for crude oil price forecasting: A complete evaluation" explores several deep learning approaches for crude oil price forecasting.

3. PROPOSED METHOD

After conducting a thorough investigation of various models to forecast daily oil prices based on historical daily oil price data, Multivariate LSTM has emerged as a promising approach. A diagram outlining the methodology used is provided in Figure 2.

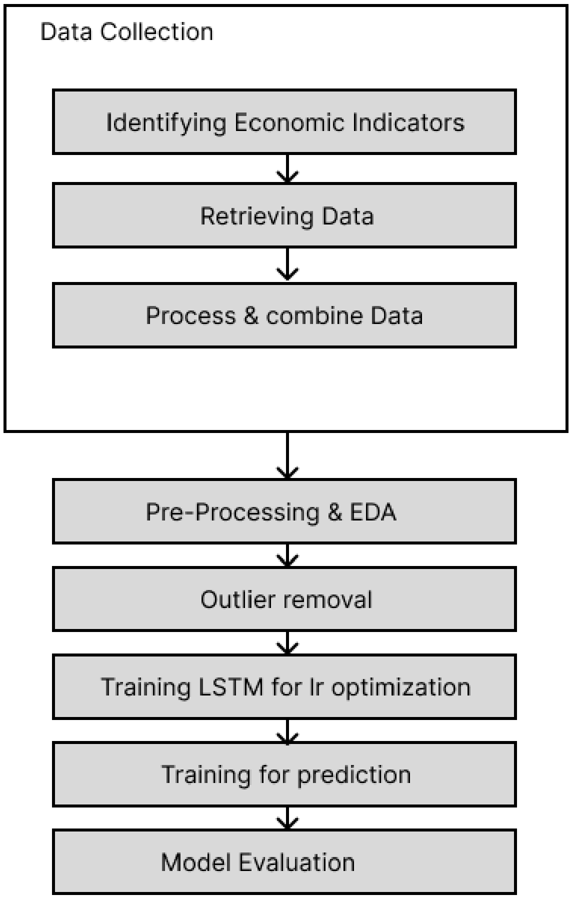

Fig-2 Method Architecture diagram

Data collection

Oil prices are affected by a variety of factors, including economic indicators such as gold, stocks, and the US dollar index **[4]**. To represent the economy and crude oil price, we have taken the historical data of the following indicators to create the dataset used for this work:

- West Texas Intermediate, or WTI [**5**], is a blend of crude oil traded on the NYMEX that includes spot prices. Together with Brent and Dubai Crude, (WTI) crude oil is a particular grade of crude oil and one of the three primary benchmarks for oil price.
- Gold Futures: When inflation rises in response to rising oil prices, so do GOLD futures. A reliable indication of economic conditions is gold futures. [4].
- US Dollar Index Futures
- Ten-Year US Bond Yield represents the yield on US 10-Year Bonds. is a certificate for the 10-year-old loan with the federal government. a good reflection of how investors feel about the economy[4].
- The S&P 500 and Dow Jones Utility Average, which measure the performance of 15 utility stocks and 500 major publicly traded corporations in the United States, respectively, using market capitalization weighting [5].

The target variable is the West Texas Intermediate (WTI) crude oil price. It is used for training because future prices are dependent on previous trends. Historical data of the target variable and other features, such as gold prices, US dollar index, and stock indicators, is collected from openly available sources for the period 2000 to 2019. The data is then combined for coinciding dates.

Data pre-processing & Exploratory Data Analysis(EDA):

The first stage of pre-processing entails the construction of a date index. Imputations from the day before are then used to replace any null values in the dataset. This is as a result of the finding that short-term changes in oil prices are often comparable. After processing, the data is next exposed to exploratory data analysis (EDA), which looks for trends, patterns, and connections between the target variable and other variables. To find outliers for each year, a box and whisker plot is used specifically, making it possible to spot data points that need to be removed. Oil prices can vary dramatically during financial crises and other big events, thus it is crucial to spot and eliminate outliers. These data points can be eliminated to improve model training.

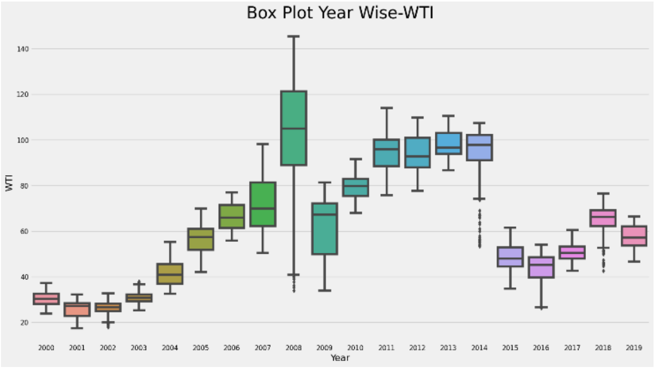

Fig 3- Box plot for year wise WTI prices

In the years 2007, 2008, and 2009, there were notable shifts in the prices of West Texas Intermediate (WTI) which can be seen in the box plot shown in Fig 3. In instance, WTI oil prices ranged between $30 and $140 in 2008, a year marked by the financial crisis and high volatility. The financial crisis that lasted from 2007 to 2009 was notable for causing significant changes in oil prices. We eliminated data points between 2007 and 2009 to lessen the influence of this crisis on our findings. This elimination procedure resulted in the removal of 482 outliers. We have also purposely left out data from 2020 and after, as that would also say a similar story due to the outbreak of COVID-19.

3.3. Model training

To avoid overfitting, we first do a train-test split before standardization. We store data using the Long Short-Term Memory (LSTM) model, which accepts a predetermined amount of time steps for each output. For instance, if we provide 60 time steps, the model will use days 1 through 60 as a sequence to forecast the results for days 61 and beyond.

We enlarge the learning rate vs. epoch graph in order to establish the ideal learning rate for our model. We choose the learning rate that results in the smallest loss, which in this instance seems to be somewhere around 3e-5. One of the most important factors in the model training process is fine-tuning the learning rate. We see that the training and validation loss curves are close to each other, demonstrating the absence of overfitting. Also, the flat loss curves are a good thing. We would have concluded that overfitting was taking place if we had seen a sizable degree of validation loss reducing with an increase in the number of epochs. Yet given that this doesn't seem to be the case, we can say that overfitting is not an issue.

Model evaluation

We assessed how well our Long Short-Term Memory (LSTM) model, which makes use of the financial outlier elimination method, performed. The following metrics [8] were used to evaluate the model:

- R2 Score: This statistic estimates the percentage of the target variable's variation that the model Accounts for. A score of 1 denotes a perfect fit, whereas a score of 0 denotes that the model is only capable of accurately predicting the target variable's mean.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (X_i - Y)^2}{\sum_{i=1}^{m} (Y - \bar{Y})^2}$$

- Mean Squared Error (MSE): This measure shows the average of the squared variations between the target variable's expected and actual values. It is a statistic that is frequently utilised in regression issues.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = Mean Squared Error
n = number of data points
Y_i = Observed Values
$\hat{Y}_i$ = Predicted Values

- Mean Absolute Error (MAE): This statistic shows the average of the absolute discrepancies between the target variable's expected and actual values. It is a frequently employed statistic in regression issues, similar to MSE.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i|$$

As the main indicators of the effectiveness of our model, we report these data. Lower MSE and MAE values imply that the model's predictions are more in line with the actual data, while higher R2 scores indicate superior model performance. To evaluate the efficacy of the employed outlier removal technique, each of these metrics is also generated for data including outliers and is compared.

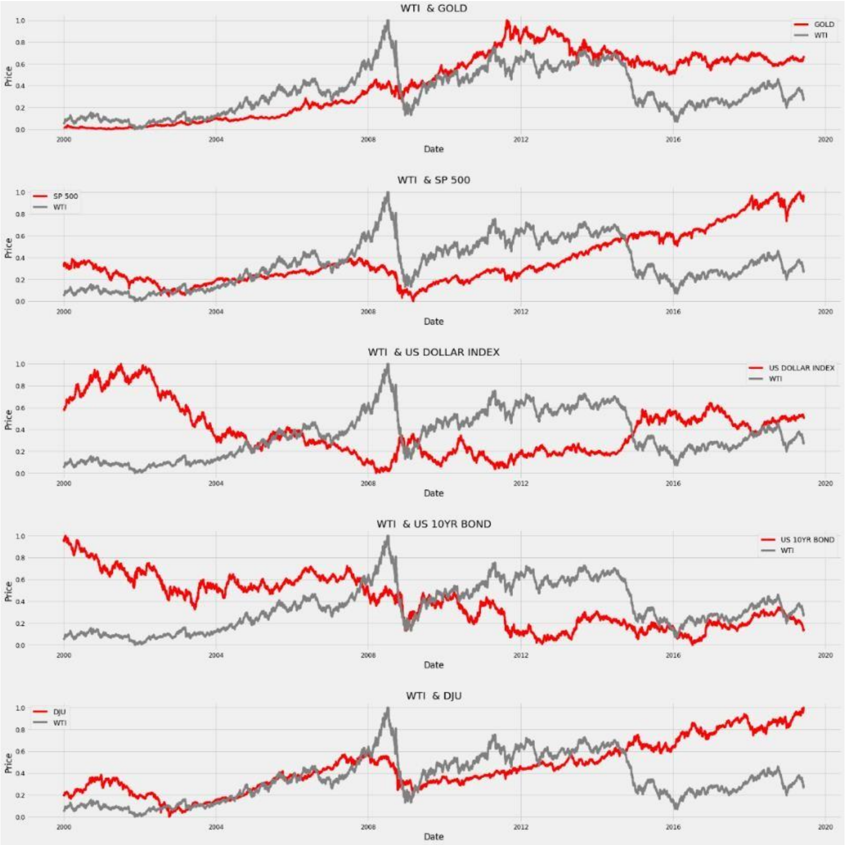

Fig- 4 Bivariate Analysis of Every Economic Indicator V/S WTI Price

4. RESULTS AND DISCUSSION

Bivariate feature analysis was used in this study to better understand the connections between various factors and how they affect WTI pricing. The findings reveal that both WTI and SP 500 saw significant declines in 2009, proving that both series were influenced by the same outside forces and moved in tandem. With the exception of rare time periods, such as 2000, 2007–09, and 2010–16, when their connection seems to be inverse, the US Dollar index and WTI have a high correlation. This implies that there may be other variables at play in addition to the US Dollar's movement having a substantial influence on the price of WTI.

Prices for WTI and gold have been seen to fluctuate in tandem, with gold occasionally appearing to lag behind WTI, especially in 2007, when there was little difference between the two values. Because that changes in US 10-year Bond interest rates can predict impending financial crises, they appear to be a leading indication of WTI price volatility. Lastly, there is a strong link between 2003 and 2009 between the movements of the DJU and WTI. Overall, these findings offer insightful information on the connections between various factors and how they affect WTI prices, which may be utilized to guide risk management and trading tactics in the oil industry.

With an R2 value of 0.937, the LSTM model that was trained utilizing the financial outlier elimination strategy demonstrated outstanding performance results, explaining 93.7% of the variation in the target variable. The mean absolute error of 2.976 and the mean squared error of 14.242 both show that the model's predictions were quite accurate. These indicators demonstrate how well the model predicts WTI prices while taking the effects of financial outliers into account and can guide risk management and investing strategies in the oil industry. The results are also visualized in the graph in Fig-5.

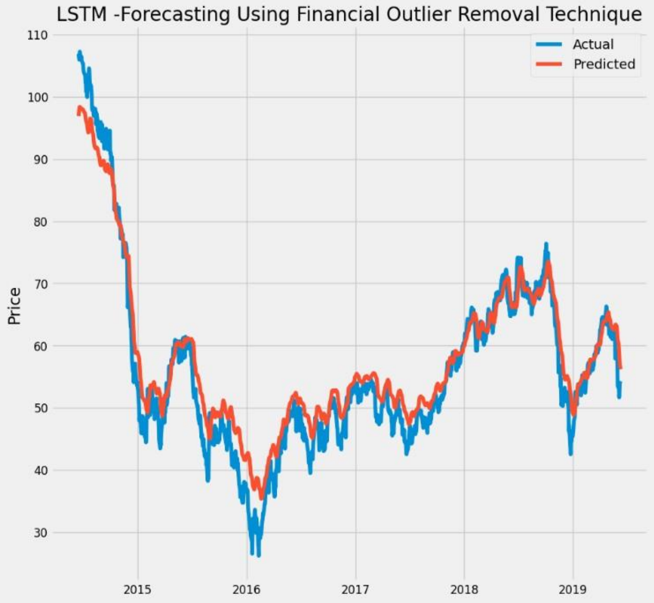

Fig-5 Predictions v/s actual for financial outlier removal

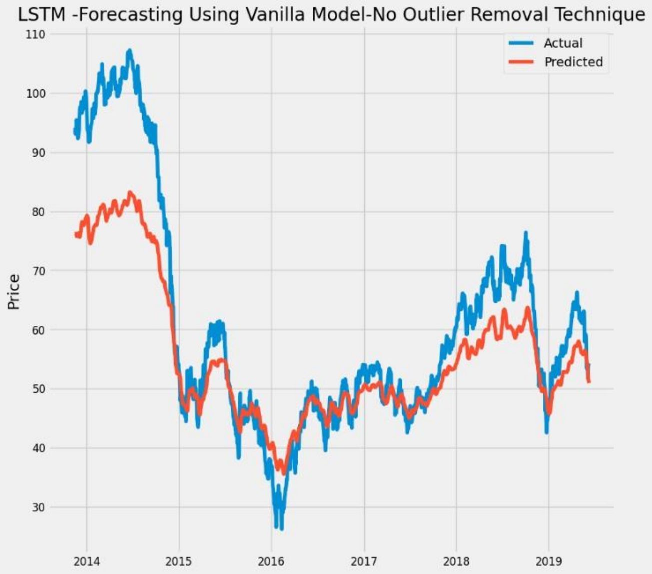

Fig-6 Predictions v/s actual for no outlier removal technique

The LSTM model trained using the financial outlier reduction method performed noticeably better than the base model without outlier removal. The outlier removal enhanced the model's capacity to explain the variance in the target variable, as seen by the R2 score of 0.937 for the outlier-removed model, which was greater than the R2 score of 0.767 for the vanilla model. Furthermore, the outlier removal enhanced the model's predictive accuracy as evidenced by the fact that the mean squared error for the outlier-removed model was 14.242 while the mean squared error for the vanilla model was 87.915. Last but not least, the outlier-removed model performed better than the vanilla model, as evidenced by the fact that its mean absolute error, which was 2.976 instead of 6.583, was lower.

	Financial outlier	**Vanilla**
R2 score	0.937	0.767
MSE	14.242	87.915
MAE	2.976	6.583

Table 1- comparing financial outlier removal technique v/s vanilla model

Overall, the LSTM model's performance in forecasting WTI prices was greatly enhanced by the outlier removal approach, demonstrating the significance of taking financial outliers into account in financial modelling and analysis.

5. CONCLUSION

We used a new method to predict WTI crude oil prices using a type of artificial intelligence called a long short-term memory (LSTM) model. We also removed any unusual data points (outliers) from our data set, which improved the accuracy of our predictions. Our model with outlier removal had an R2 score of 0.937, Mean Squared Error of 14.242, and Mean Absolute Error of 2.976. This was better than the model without outlier removal, which had an R2 score of 0.767, Mean Squared Error of 87.915, and Mean Absolute Error of 6.583.

We used a multivariate approach by including various economic indicators as inputs to our model. This allowed us to capture complex relationships and dependencies between different factors, which improved the accuracy and robustness of our predictions.

Our study highlights the importance of considering multiple factors when predicting financial outcomes. It also demonstrates the effectiveness of LSTM models in such scenarios. Overall, our research provides valuable insights into the predictive modeling of WTI crude oil prices and could potentially be used to improve financial decision-making.

6. REFERENCES

1. N. Gupta and S. Nigam, "Crude Oil Price Prediction using Artificial Neural Network," *Procedia Computer Science*, vol. 170, pp. 642–647, Jan. 2020, doi: 10.1016/j.procs.2020.03.136.
2. Z. Cen and J. Wang, "Crude oil price prediction model with long short term memory deep learning based on prior knowledge data

transfer," *Energy*, vol. 169, pp. 160–171, Feb. 2019, doi: 10.1016/j.energy.2018.12.016 .

3. J. Ju and F. Liu, "Multivariate Time Series Data Prediction Based on ATT-LSTM Network," *Applied Sciences*, vol. 11, no. 20, p. 9373, Oct. 2021, doi: 10.3390/app11209373.
4. M. Arfaoui and A. B. Rejeb, "Oil, gold, US dollar and stock market interdependencies: a global analytical insight," *European Journal of Management and Business Economics*, vol. 26, no. 3, pp. 278–293, Oct. 2017, doi: 10.1108/ejmbe-10-2017-016.
5. H. Geman and C. Kharoubi, "WTI crude oil Futures in portfolio diversification: The time-to-maturity effect," *Journal of Banking and Finance*, vol. 32, no. 12, pp. 2553–2559, Dec. 2008, doi: 10.1016/j.jbankfin.2008.04.002 .
6. H. Chiroma, S. Abdulkareem, and T. Herawan, "Evolutionary Neural Network model for West Texas Intermediate crude oil price prediction," *Applied Energy*, vol. 142, pp. 266–273, Mar. 2015, doi: 10.1016/j.apenergy.2014.12.045 .
7. O. Orojo, J. Tepper, T. McGinnity, and M. Mahmud, *A Multi-recurrent Network for Crude Oil Price Prediction*. 2019. doi: 10.1109/ssci44817.2019.9002841.
8. Prabowo, H., Hidayat, A. A., Cenggoro, T. W., Rahutomo, R., Purwandari, K., & Pardamean, B. (2021). *Aggregating Time Series and Tabular Data in Deep Learning Model for University Students' GPA Prediction. IEEE Access, 9, 87370–87377.* doi:10.1109/access.2021.3088152