E-ISSN : 2963-4946
Vol. 2 No. February 5, 2024
Asian Journal of Social and Humanities
https://ajosh.org/

Comparative Performance of Learning Methods In Stock Price Prediction Case Study: MNC Corporation

Rifqi Khairurrahman, Gerry Firmansyah, Budi Tjahjono, Agung Mulyo Widodo
Universitas Esa Unggul, Indonesia
E-mail: eaglemaster7@gmail.com, gerry@esaunggul.ac.id, agung.mulyo@esaunggul.ac.id, budi.tjahjono@esaunggul.ac.id

***Correspondence:** eaglemaster7@gmail.com

KEYWORDS
stock prediction; CNN; RNN; LSTM; GRU; MLP

ABSTRACT
Shares are a popular business investment, the development of information technology now allows everyone to buy and sell shares easily online, investment players, both retail and corporate, are trying to make predictions. The purpose of this study is to find out comparative performance of learning methods in stock price prediction. There are currently many research papers discussing stock predictions. using machine learning / deep learning / neural networks, in this research the author will compare several superior methods found in the latest paper findings, including CNN, RNN LSTM, MLP, GRU and their variants. From the 16 result relationships and patterns that occur in each variable and each variable is proven to show its respective role with its own weight, in general we will summarize the conclusions in chapter V below, but in each analysis there are secondary conclusions that we can get in detail. The variable that has the most significant effect on RMSE is variable B (repeatable data) compared to other variables because it has a difference in polarity that is so far between yes and no. The configuration of input timestep (history)=7 days and output timetep (prediction)=1 day is best for the average model in general.

Introduction

Prediction of stock price movements is the subject of many studies that are trending today. On the one hand, we have proponents of the efficient Market Hypothesis who claim that stock prices are unpredictable. Research has shown that, if modeled correctly, stock prices can be predicted with quite the degree of reasonable accuracy that can be obtained with fundamental analysis and technical analysis (Ahire et al., 2021).

Stocks are business investments in both large investor classes and retail investors. Business processes in the stock exchange run between brokers who make stock buying / selling transactions. Consideration Buying and selling shares is done by looking at the company's prospects. Before making stock transactions, business people conduct analysis

by looking at stock price charts to determine timing and action whether it is best to buy sell or hold / wait (Ahmad et al., 2022).

CNN, Deep learning and other Artificial intelligence methods are the final frontier of science in the computer field today, there has been a lot of AI literature that reviews the benefits and development with this method. On the other hand, today's online world makes stock trading easier and closer to everyone, the amount of capital and transaction volume is also getting bigger. From this the general public and also brokers are felt to need an assistant system to help estimate predictions with a reasonable margin of error (Bibi et al., 2020).

The purpose of this study was to find out which method is best from several machine learning / deep learning in predicting 3 MNC stocks in multiple time windows.

Research Methods

Test Variables

Variable A : ML / DL Model

There are 8 models selected for testing including the following:

- LSTM (Vanilla/ standard) Simple method 1 layer LSTM
- LSTM (multi stack) LSTM method that is multi layer
- LSTM (Bidirectional) Alternating LSTM method
- CNN (Convolutional neural network)
- RNN (Recurrent neural network)
- MLP (Multilayer perceptron)
- GRU (simple Gated recurrent unit)
- GRU (multilayer Gated recurrent unit)

Variable B: data repetitiveness

Data repetitivienes are the second variable Boolean value where if the variable is true (True) where training data and testing data use repeated source data, the looping length of the data corresponds to variables C and D, for example:

If we have a data source S=[1,2,3,4,5,6,7], and if we use a loop with timestep 3 input and timestep 2 output, then the data becomes X = [[1,2,3], [2,3,4],[3,4,5]] and Y=[[4,5], [5,6],[6,7]] (Deng & Yu, 2014)

2nd example if we use input timestep =5 and output timestep=1 then it becomes X=[[1,2,3,4,5],[2,3,4,5,6]] and Y=[[6,7]] and so on according to the corresponding variables C and D ...

Variable C: input timestep / history

This variable is the amount of sequence data used (in this study it is days) input timestep=3 means using the previous 3 days history to be entered into the model

Variable D: output timestep / prediction

Likewise, with this variable this variable is the output of the prediction result, the output timestep = 5 means to produce an array sequence output with 5 data

All tested in the presence of 3 stocks namely MNCN, KPIG and BHIT

Furthermore to ensure this experiment is fair each model is pure with a normal number of units (100 units), and 200 Epoch 32 Batches, there is no special model/tweaking combination

Rules in model designing

To ensure that all is fair, the author specifies the following rules:

- The eight models use the same callback, which uses the same earlystopping callback, namely mode = auto, monitor = val_loss (validation) and patience = 5, although the

number of epoch training that will be carried out later will be different automatically following the results of different validation_loss for each configuration
- The eight models were compiled using the same method with the same optimizer, the ADAM optimizer, and loss calculation = MSE
- The eight models have the same number of layers simple / Single model strinct using only 1 layer, while for multiple using 2 layers
- All models use the same number of units i.e. 128 units, either on single or multilayer, and all

Tools

In this study the tools used are as follows:
- Python 3.8 environment in PyCharm IDE 2023 community edition
- Tensorflow & SKLearn
- Matlab &; Ms Excel
- Windows 10 x64 Xeon E2960 v3 / Ram 8 GB.

Results and Discussions

Variable analysis A (Model)

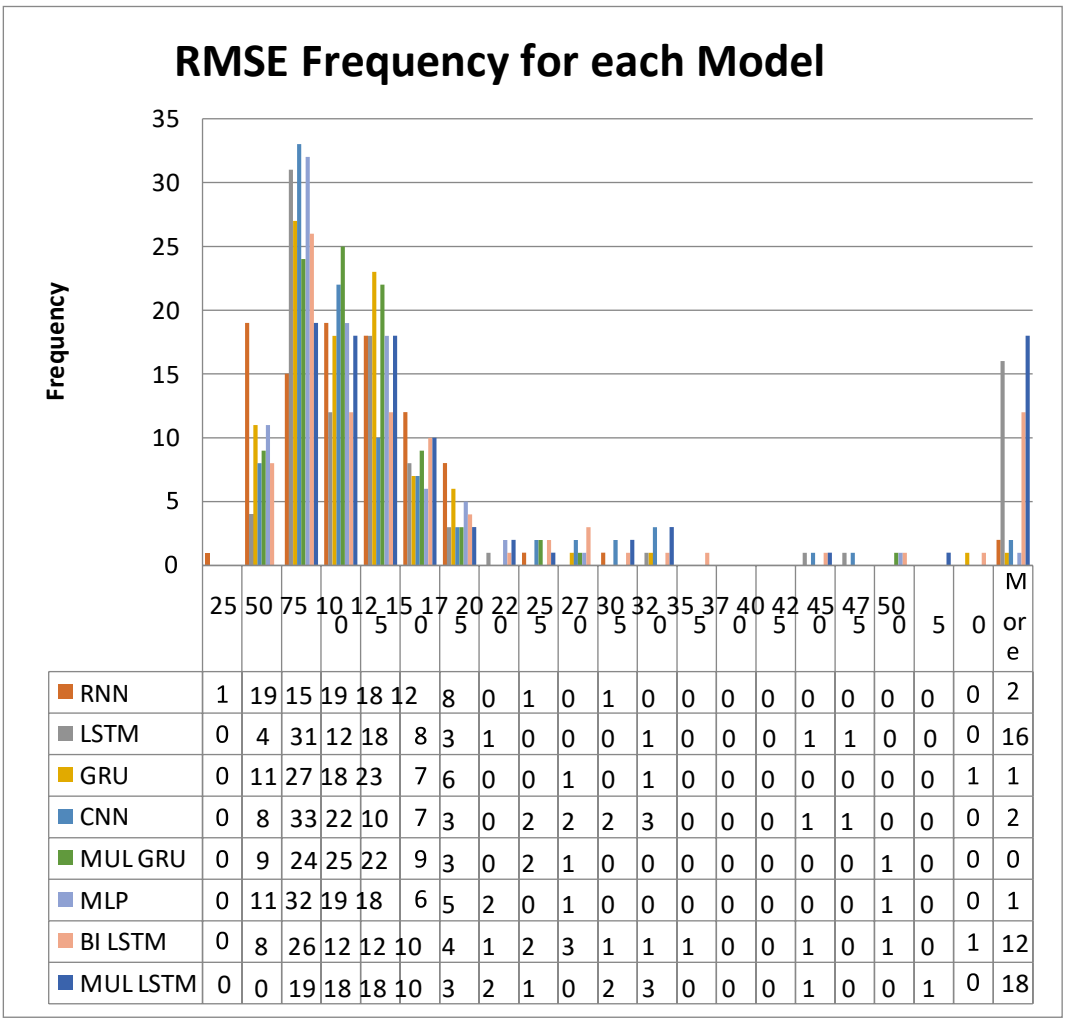
The following is the result of a comparison of ML models in the average RMSE for each combination, the RMSE result in each model is the average of 96 kinds of combinations with a total of 96x8 = 760 overall predictions, for the record of each anga in the RMSE the smaller the value the better the prediction (Trivedi & Patel, 2022).

TYPE	AVERAGE	MIN	MAX	
RNN	28,86458333	6	81	
LSTM	24,33333333	8	83	
GRU	26,21875	8	72	
CNN	23,47916667	6	65	
MGRU	23,75	7	68	
MLP	27,23958333	6	62	
BILSTM	22,03125	8	68	
MLSTM	19,9375	7	62	EPOCH
RNN	102,9550177	20,16031395	627,2010514	
LSTM	501,6625341	39,16586256	5649,901995	
GRU	102,6963024	32,14678652	634,8314789	
CNN	117,6407164	41,85433257	783,5996789	
MGRU	96,59055862	28,742268	426,1534032	
MLP	101,5619685	33,27610395	1007,12436	
BILSTM	329,4193142	39,37646458	4070,844533	
MLSTM	718,7703617	53,26609904	7773,578637	RMSE
RNN	0,012102975	0,002302176	0,093346924	
LSTM	0,064829054	0,004895594	0,795196955	
GRU	0,01181979	0,003669882	0,093873416	
CNN	0,014684713	0,00484417	0,111183502	
MGRU	0,01103379	0,003239099	0,057462184	
MLP	0,012221358	0,0038192	0,152237475	
BILSTM	0,040898438	0,004608999	0,510670851	

MLSTM	0,091125646	0,005912237	1,096643576	MAPE

From the table above, it can be seen that multiple GRU is slightly superior to other models, for Multilayer epoch training LSTM has the least epoch average of 19.9 times, while the slight epoch results are also accompanied by the highest error rate of 0.09 (Eapen et al., 2019).

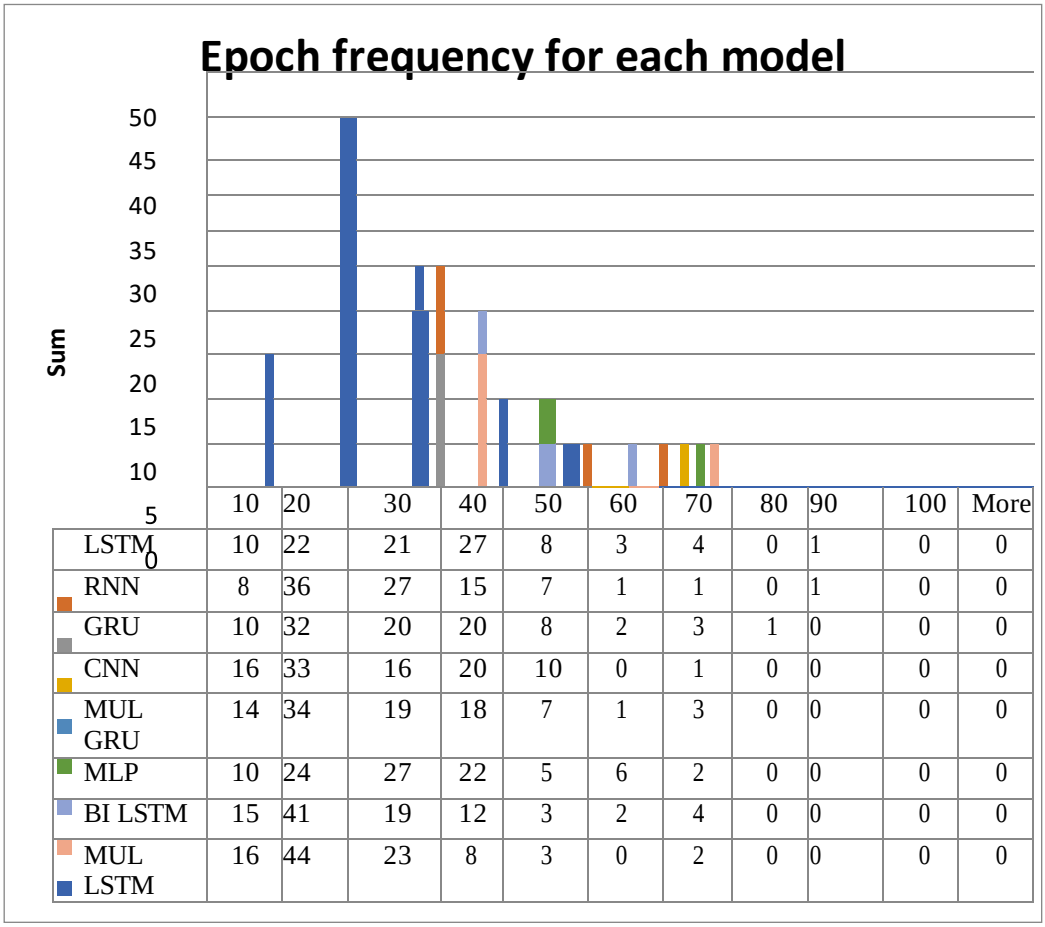
Analysis of the RMSE distribution on variable A (Model)



	25	50	75	100	125	150	175	200	225	250	275	300	325	350	375	400	425	450	475	500	5	0	More
RNN	1	19	15	19	18	12	8	0	1	0	1	0	0	0	0	0	0	0	0	0			2
LSTM	0	4	31	12	18	8	3	1	0	0	0	1	0	0	0	1	1	0	0	0			16
GRU	0	11	27	18	23	7	6	0	0	1	0	1	0	0	0	0	0	0	0	1			1
CNN	0	8	33	22	10	7	3	0	2	2	2	3	0	0	0	1	1	0	0	0			2
MUL GRU	0	9	24	25	22	9	3	0	2	1	0	0	0	0	0	0	0	1	0	0			0
MLP	0	11	32	19	18	6	5	2	0	1	0	0	0	0	0	0	0	1	0	0			1
BI LSTM	0	8	26	12	12	10	4	1	2	3	1	1	1	0	0	1	0	1	0	1			12
MUL LSTM	0	0	19	18	18	10	3	2	1	0	2	3	0	0	0	1	0	0	1	0			18

It can be seen that the majority range of all RMSE results is mostly in the range of 50-75 classes and if you look at the shape of the curve at first glance naturally follows the normal distribution of the bell curve with a left tendency because there is also a lot of outiler / data noise where the RMSE is more than 500 and it could be a failure of some models in testing, but in general all models managed to predict well (Rasyid et al., 2021).

Training Epoch analysis on variable A (model)



	10	20	30	40	50	60	70	80	90	100	More
LSTM	10	22	21	27	8	3	4	0	1	0	0
RNN	8	36	27	15	7	1	1	0	1	0	0
GRU	10	32	20	20	8	2	3	1	0	0	0
CNN	16	33	16	20	10	0	1	0	0	0	0
MUL GRU	14	34	19	18	7	1	3	0	0	0	0
MLP	10	24	27	22	5	6	2	0	0	0	0
BI LSTM	15	41	19	12	3	2	4	0	0	0	0
MUL LSTM	16	44	23	8	3	0	2	0	0	0	0

From the results of the figure in the duration of training epoch over training occurs in the 20s, and the graph if you look at it naturally already follows the normal distribution of the bell curve, meaning that the graph data that we see reliably already reflects a natural representation, and there is no outlier epoch on this variable (Caniago et al., 2021).

Anomaly in LSTM

From the above points previously it is said that it is generally seen that all LSTM variants have RMSE, especially multilayer LSTM (which is marked by a ceiling far above other models 5-6 times adrift can be seen from the diaghram below, why is it so much? Because it was found that LSTM's predictive capabilities were far missed, worse in processing non-repeatable data (Khan et al., 2021).

If we examine the difference between repeatable vs non-repeatable for LSTM we find the following results:

Type	Repeat	RMSE
Simple LSTM	Yes	114.7063266
	No	1007.798039
Multilayer LSTM	Yes	93.6343394
	No	1280.36542
Bidirectional LSTM	Yes	76.9121999

	No	930.1901084
Multilayer GRU	Yes	87.00819945
	No	110.5796003

The difference in variable B between repeat data vs non-repeat data is almost 10 x for all LSTM variants, compared to multilayer GRU which is very little difference, LSTM is very affected by this variable, there is also an interesting thing in this case that we can observe, in the case of repeatable bidirectional data LSTM has better performance outperforming multilayer GRU (Shukla et al., 2022).

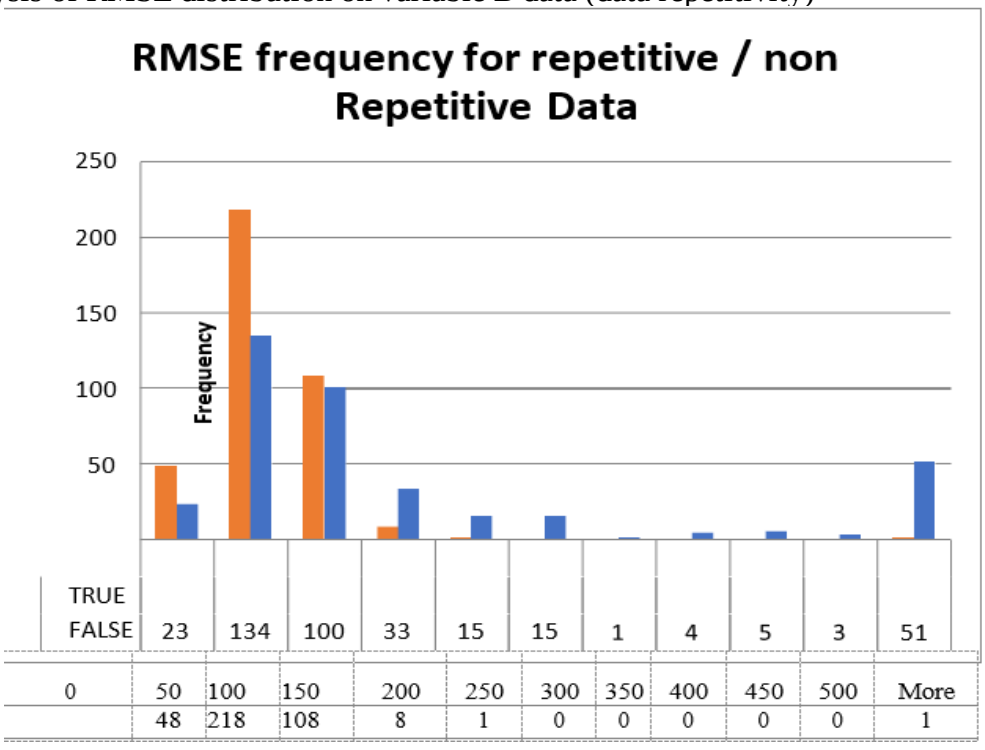
Variable B Analysis (Repetitiveness)

Not only LSTM is greatly affected by data reptitivenes, all models are actually all affected by this factor and on average for non-repeatable data has worse results of 5x or more (Ludwig, 2019), here is a repeatable-data comparison for all models:

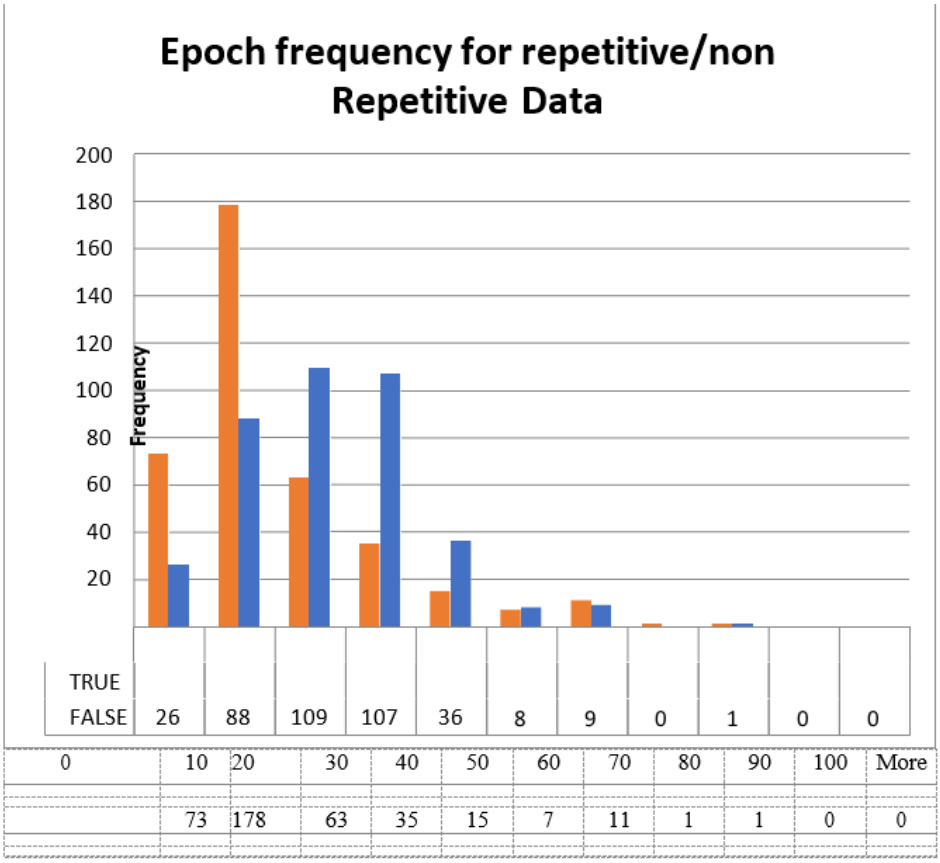
REPEAT	Average	MIN	MAX	UNIT
TRUE	20,72135417	7	81	
FALSE	28,2421875	6	83	EPOCH
TRUE	88,0670136	20,16031395	2106,922316	
FALSE	429,7571798	35,52635452	7773,578637	RMSE
TRUE	0,010088418	0,002302176	0,301586422	
FALSE	0,054590524	0,004291345	1,096643576	MAPE

From the data above we see that between repeat data and non-repeat has a very very wide difference, why is that? The author has 2 further hypotheses about this, first repeat data has more data feeds and richer richer so that variations in matches about predictions are more biased captured by ML / DL, the second factor is the small possibility of data overfit due to overflow by history from previous training especially for models that use memory history such as LSTM (Mishra et al., 2021).

Analysis of RMSE distribution on variable B data (data repetitivity)



From the image above, we observe that for TRUE, it is always more on the left (orange bar higher) in the low RMSE data range than FALSE (blue bar) which tends to be distributed to the right (higher RMSE range) there is a clear difference in distribution, especially for FALSE whose outliers are above 50, so it can be concluded unequivocally that TRUE has better performance (Nabipour et al., 2020).
Training epoch analysis on variable B (Repetitivenes)



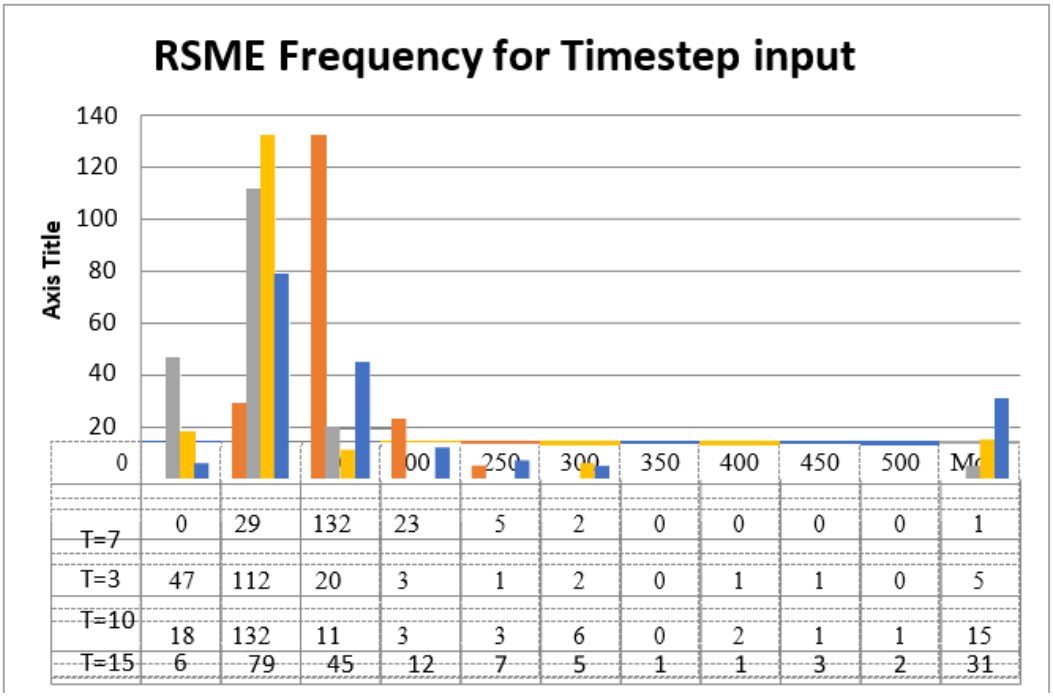
From the Figure above you can see the difference again, TRUE (orange bar has a median point in the 20s and tapered while False has a more even median point between 30-40 which shows the epoch for TRUE is more convergent, and False is more divergent and the book of Jesus concludes that True has less epoch training (Namdari & Durrani, 2021).
Variable Analysis C (input timestep)

T INPUT	**AVERAGE**	**MIN**	**MAX**	**UNIT**
t=3	20,76041667	8	68	EPOCH
t=7	25,61458333	7	81	
t=10	24,34375	6	66	
t=15	27,20833333	6	83	
t=3	127,4795499	87,60711516	697,3507006	RMSE
t=7	102,4603947	20,16031395	1226,766746	
t=10	180,0123783	32,03995796	2521,770851	
t=15	625,696064	36,78043449	7773,578637	
t=3	0,014264137	0,009727645	0,09584945	

t=7	0,0121487	0,002302176	0,169293866	MAPE
t=10	0,022843166	0,003637193	0,327559112	
t=15	0,08010188	0,004149383	1,096643576	

From the table above, it can be concluded that T input = 7 is the best and T input = 15 is the worst, but the training epoch T = 3 is the least, there is an interesting thing that the results of T = 3 .7 10 , and 15 results are not linear where we cannot say that the higher the input T the worse it is, however, there is a certain point where the correct input T value will produce more optimal predictions and thereafter continue to deteriorate as the T value increases (Hoseinzade & Haratizadeh, 2019).

Analysis of RMSE distribution on variable C (input timestep)

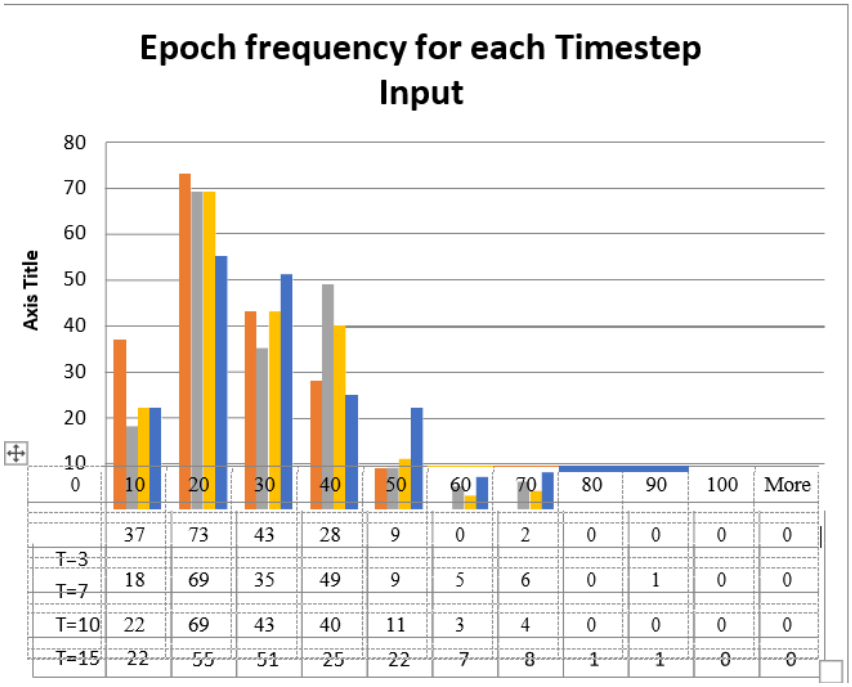


	0	29	132	23	5	2	0	0	0	0	1
T=7											
T=3	47	112	20	3	1	2	0	1	1	0	5
T=10	18	132	11	3	3	6	0	2	1	1	15
T=15	6	79	45	12	7	5	1	1	3	2	31

From the Figure above we see that there are that for T=10 and T=15 there are many outliers where the RMSE result of more than 500 is what makes T=10 lose to T=7 although it can be seen that T=10 wins in the 100s and T=7 wins in the 150s, besides that in general we can see that the curve has been seen following a normal distribution where we can say this is quite representative with angles The view we observe

The author also has a sub-hypothesis that a lot of input T will also require longer training and tuning to be able to predict well, therefore T that exceeds the optimal limit will lose.

Dristribustion Analysis Epoch training for Variable C (input timestep)



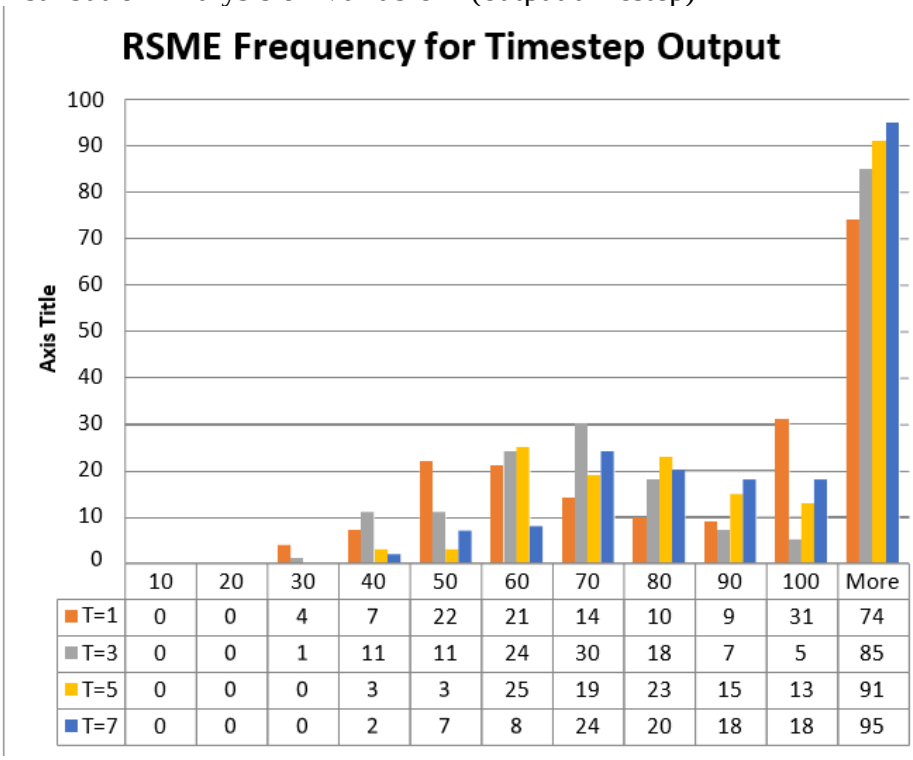
From the figure above we can see that the majority of training epochs are in the 20s, there is not much difference between the four data, and after the peak point in the 20s and the higher the epoch the fewer the number in this test, the graph also shows the normal distribution shape panning to the left.

Variable D Analysis (output timestep)

T OUTPUT	AVERAGE	MIN	MAX	UNIT
T_OUT=1	21,68229167	6	81	EPOCH
T_OUT=3	24,09375	6	69	
T_OUT=5	25,93229167	8	83	
T_OUT=7	26,21875	8	70	
T_OUT=1	232,2241972	20,16031395	4608,030448	RMSE
T_OUT=3	312,6007879	28,94929418	7773,578637	
T_OUT=5	235,9101403	36,61536638	4755,851117	
T_OUT=7	254,9132613	32,03995796	5355,351469	
T_OUT=1	0,028852344	0,002302176	0,536735398	MAPE
T_OUT=3	0,040810064	0,003239099	1,096643576	
T_OUT=5	0,028481104	0,004212775	0,620125476	
T_OUT=7	0,03121437	0,003637193	0,757823378	

This is quite an interesting / strange variable, judging from this table there is actually a high inequality where the maximum RMSE is in the thousands while the average (average) of all these variables in the 200s means there are outliers with little data but have a very high RMSE. This requires further research that separates why it happens and what variables influence, but it can be said that in this point T out = 1 is superior but narrowly adrift with Tout = 3.

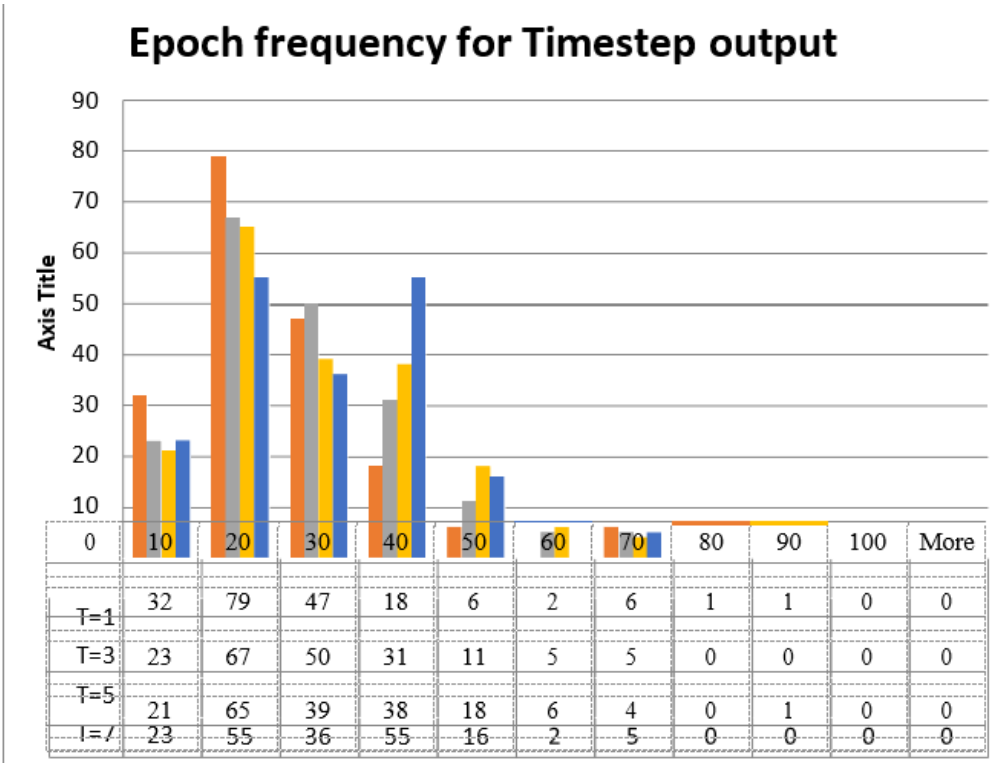
RMSE Distribution Analysis on Variable D (output timestep)



	10	20	30	40	50	60	70	80	90	100	More
T=1	0	0	4	7	22	21	14	10	9	31	74
T=3	0	0	1	11	11	24	30	18	7	5	85
T=5	0	0	0	3	3	25	19	23	15	13	91
T=7	0	0	0	2	7	8	24	20	18	18	95

It can be seen directly that, the RMSE outlier of more than 100 is quite high, and that might affect the mean RMSE on this variable, where if we look back RMSE actually peaked in the 70s and continued to slope upwards the phenomenon needs further research on this subject.

Analysis of Epoch Distribution on Variable D (output timestep)



	10	20	30	40	50	60	70	80	90	100	More
T=1	32	79	47	18	6	2	6	1	1	0	0
T=3	23	67	50	31	11	5	5	0	0	0	0
T=5	21	65	39	38	18	6	4	0	1	0	0
T=7	23	55	36	55	16	2	5	0	0	0	0

The epoch peaks in the 20s class range and the higher the epoch the fewer the number, the least epoch is T=1 and the highest epoch is T=7 this graph also follows the normal distribution of the bell curve which is already represented naturally.

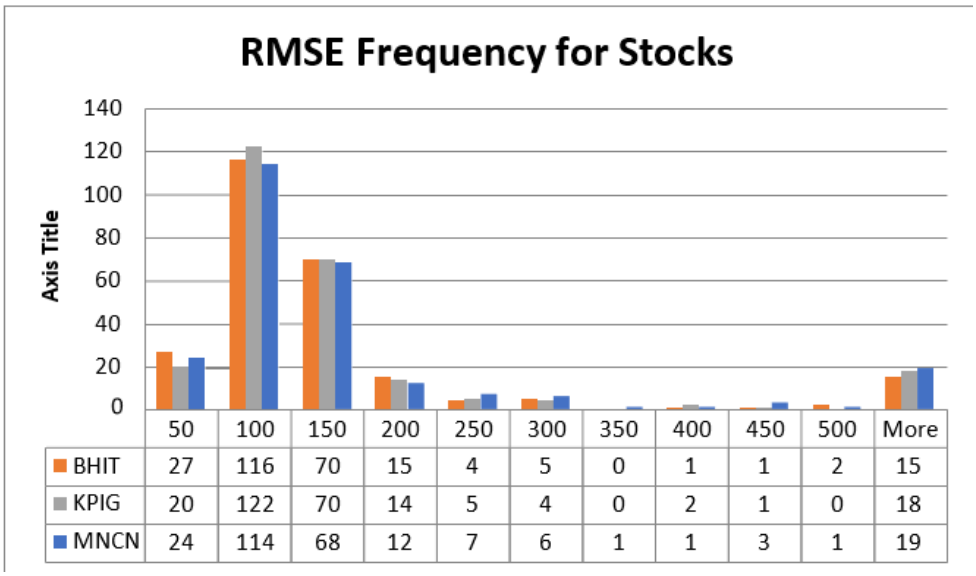
Variable Analysis D (Stocks)

Actually this is a comparison variable, but we will analyze it too, here is the data:

STOCK	AVERAGE	MIN	MAX	UNIT
BHIT	24,61328125	6	83	
KPIG	24,5546875	6	81	
MNCN	24,27734375	6	69	EPOCH
BHIT	251,9673095	29,66951721	7773,578637	
KPIG	246,405199	20,16031395	5714,7373	
MNCN	278,3637816	28,62764747	5547,175371	RMSE
BHIT	0,030746159	0,003571725	1,096643576	
KPIG	0,031044407	0,002302176	0,795196955	
MNCN	0,035227846	0,003239099	0,807824281	MAPE

This variable is also interesting and there is a strangeness, there is a difference between RMSE and MAPE, in the test results, usually RMSE is also along with MAPE but if the data results are very similar then MAPE can also be different from RMSE depending on how we look at the data, the thing that can be concluded from this variable is that all models succeed in predicting and have similar results.

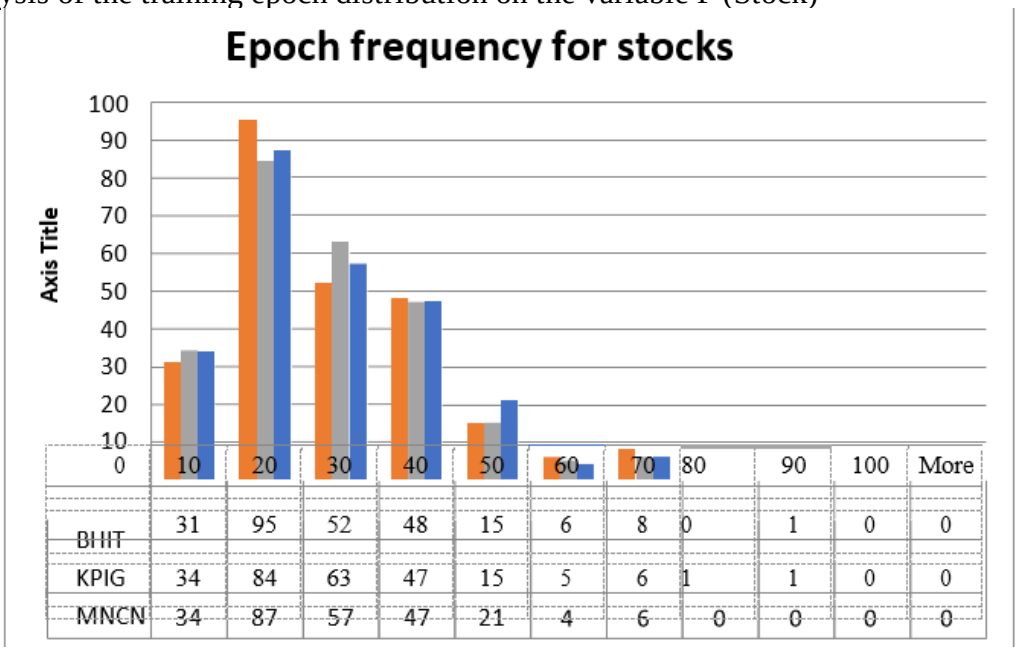
RMSE variable distribution analysis F (Shares)



	50	100	150	200	250	300	350	400	450	500	More
BHIT	27	116	70	15	4	5	0	1	1	2	15
KPIG	20	122	70	14	5	4	0	2	1	0	18
MNCN	24	114	68	12	7	6	1	1	3	1	19

It can be seen that there are outliers for the RMSE above 500 that affect the average / mean RMSE, but graphically you can see the RMSE peaks in the 100s, there is not much difference for the RMSE on this variable, nothing stands out

Analysis of the training epoch distribution on the variable F (Stock)



	10	20	30	40	50	60	70	80	90	100	More
BHIT	31	95	52	48	15	6	8	0	1	0	0
KPIG	34	84	63	47	15	5	6	1	1	0	0
MNCN	34	87	57	47	21	4	6	0	0	0	0

From the Figure above, it can be concluded that the most epochs are found in the 20s class range, then the higher the epoch, the number decreases, the smallest MNCN with an average of 24.2 but also has the worst RMSE results, in this variable the data are similar and there is not much difference in epochs in this variable.

From the 16 results of the analysis above there are relationships and patterns that occur in each variable and each variable is proven to show its respective role with its own weight, in general we will summarize the conclusions in chapter V below, but in each analysis there are secondary conclusions that we can get in detail.

Please note, the results of this study are very concerned with the amount of data or samples used, the more statistical data the more credible it is. The total number of experiments worked on above is (Variable A/Model=8) X (Variable B/Repetitiveness=2) X (Variable C: input timestep=4) X (Variable D: output timestep=4) X (Variable D: stock=3) = 8 x 4 x 4 x 3 = 768 experimental combinations, all experiments are carried out in about 2 days by a batch program (with specifications in chapter III) whose detailed data results are in the appendix.

Conclusion

Here are the conclusions in this study: multilayer GRU has the best performance (with an average RMSE of 96.5) among the eight methods compared but by a slight margin with other models

The variable that has the most significant effect on RMSE is variable B (repeatable data) compared to other variables because it has a difference in polarity that is so far between yes and no

The configuration of input timestep (history)=7 days and output timetep (prediction)=1 day is best for the average model in general.

The three types of MNC stocks are successfully predicted well by the eight models in general, when referring to the RMSE 300 threshold, even though the three stocks have different characteristics and trends

References

Ahire, P., Lad, H., Parekh, S., & Kabrawala, S. (2021). LSTM based stock price prediction. *International Journal of Creative Research Thoughts*, *9*(2), 5118–5122.

Ahmad, G. I., Singla, J., Anis, A., Reshi, A. A., & Salameh, A. A. (2022). Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus-A Comprehensive Review. *International Journal of Advanced Computer Science and Applications*, *13*(2).

Bibi, I., Akhunzada, A., Malik, J., Iqbal, J., Musaddiq, A., & Kim, S. (2020). A dynamic DL-driven architecture to combat sophisticated Android malware. *IEEE Access*, *8*, 129600–129612.

Caniago, A. I., Kaswidjanti, W., & Juwairiah, J. (2021). Recurrent Neural Network With Gate Recurrent Unit For Stock Price Prediction. *Telematika: Jurnal Informatika dan Teknologi Informasi*, *18*(3), 345–360.

Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends® in signal processing*, *7*(3–4), 197–387.

Eapen, J., Bein, D., & Verma, A. (2019). Novel deep learning model with CNN and bi-directional LSTM for improved stock market index prediction. *2019 IEEE 9th annual computing and communication workshop and conference (CCWC)*, 264–270.

Hoseinzade, E., & Haratizadeh, S. (2019). CNNpred: CNN-based stock market prediction using a diverse set of variables. *Expert Systems with Applications*, *129*, 273–285.

Khan, M., Wang, H., Riaz, A., Elfatyany, A., & Karim, S. (2021). Bidirectional LSTM-RNN-based hybrid deep learning frameworks for univariate time series classification. *The Journal of Supercomputing*, *77*, 7021–7045.

Ludwig, S. A. (2019). Comparison of time series approaches applied to greenhouse gas analysis: ANFIS, RNN, and LSTM. *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6.

Mishra, R. K., Reddy, G. Y. S., & Pathak, H. (2021). The understanding of deep learning: a comprehensive review. *Mathematical Problems in Engineering*, *2021*, 1–15.

Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., & Salwana, E. (2020). Deep learning for stock market prediction. *Entropy*, *22*(8), 840.

Namdari, A., & Durrani, T. S. (2021). A multilayer feedforward perceptron model in neural networks for predicting stock market short-term trends. *operations research forum*, *2*(3), 38.

Rasyid, A. F., Agushinta, D., & Ediraras, D. T. (2021). Deep Learning Methods In Predicting Indonesia Composite Stock Price Index (IHSG). *International Journal of Computer and Information Technology (2279-0764)*, *10*(5).

Shukla, S. K., Joshi, K., Singh, G. D., & Dumka, A. (2022). Stock Market Prediction Using Deep Learning. *2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP)*, 91–95.

Trivedi, D. V., & Patel, S. (2022). An Analysis of GRU-LSTM Hybrid Deep Learning Models for Stock Price Prediction. *International Journal of Scientific Research in Science, Engineering and Technology*, *9*(3), 47–52.