A Comparative Study of Data Augmentation Techniques in Stock Return Forecasting

Bijaya Kumar Nanda,Girija Prasan Das,Fernadis Mishra
Dept. of Computer Science and Engineering,Gandhi Institute For Technology,Bhubaneswar,752054
Email:girija@gift.edu.in

ABSTRACT

In the last decade, there have been advances in machine learning performance in various domains, including image classification, natural language processing, and speech recognition. The increase in the size of training data is essential for the improvement in these domains. The two ways to have larger training sets are acquiring more original data and employing effective data augmentation techniques. However, in stock prediction studies, the sizes of datasets have not changed much and there is no accepted data augmentation technique. Consequently, there has been no similar progress in stock prediction. This paper proposes an intuitive and effective data augmentation technique for stock return prediction. New synthetic stocks are generated from linear combinations of original stocks. Unlike previous studies, our augmentation mimics actual financial asset creation processes. Our data augmentation significantly improves prediction accuracy. Moreover, we investigate how the characteristics of original data affect the data augmentation performance. We find a U-shape relationship between accuracy improved from the augmentation and return correlation in original data.

1. INTRODUCTION

Recently, machine learning (ML) and artificial intelligence (AI) advance quickly in various domains, including text, image, speech and games. In these domains, machines perform as well as or better than humans [1]–[6]. The increase in the size of training data is a key for this improvement. The two ways to have larger training sets are i) acquiring new original data or ii) employing data augmentation techniques. Data augmentation is a method to increase the amount of data for training mo dels without collecting new data. For example, in the case of image processing, a photo of an orange could be used to generate thousands of different orange images by various image processing techniques such as rotating, flipping, and blurring. Data augmentation has been proven fruitful in improving model performance [7].

However, there is no analogous progress in prediction performance in the studies of stock prediction. For example, economists could not predict the S&P 500's 40 percent plunge in March 2020 and a subsequent new high in August 2020. Unlike other domains, the age of big data does not significantly increase the size of training sets in stock prediction studies. In other domains such as image classification, the number of images available for training increased exponentially in recent years. Nevertheless, the number of

stock series is relatively the same. Moreover, well-accepted data augmentation techniques for stock price series do not exist [8], [9]. Few data augmentation techniques for stock data have been proposed. Existing studies apply data augmentation techniques based on signal processing to stock price prediction [10], [11]. However, stock prices are different from physical waves. Therefore, stock data augmentation based on signal processing has no solid economic foundation. To our knowledge, this study is the first data augmentation study on stock return prediction with a solid economic foundation. Unlike existing studies, our data augmentation is sensible and mimics actual financial asset creation. New augmented assets and price series are generated from a linear combination of existing assets/stocks.

The paper is organized as section 1 explains the data augmentation process. Section 2 proposes a data augmentation technique. Section 3 and 4 apply the proposed data augmentation for a stock return prediction and show that the data augmentation significantly improves prediction performance. In section 5, we study how the characteristics of original data affect the performance of data augmentation. The final section concludes.

2. DATA AUGMENTATION METHOD
2.1. Generating new assets

Table 1 depicts well-accepted data augmentation employed in various domains. In these domains, data augmentation is mostly based on invariant transformation. For example, image generating processes such as rotation and reflection are used for image augmentation. In natural language processing (NLP), back translation and synonyms are used to enrich the training data and enhance training performance. In speech recognition and time-series classification, time warping is used for data augmentation. However, to our knowledge, in existing studies, there is no well-accepted data generating process for stock data augmentation. We propose a sensible stock data augmentation mimicking actual financial asset creation processes.

Table 1. Well accepted data augmentation in various domains

Data type/domains	Data augmentation process	Studies
Image	Rotation, blurring, reflection	[12], [13]
Text	Back translation, synonym	[14], [15]
Speech and time series	Time warping	[16]–[19]
Stocks	-	-

Data augmentation for stocks generates new stocks/assets and price series from existing stocks. In financial asset management, the most common way to create a new asset is by combining underlying assets with fixed weights. For example, well-known assets constructed by weighting underlying assets are S&P 500 index funds. An S&P 500 index fund is market-capitalization-weighted of 505 U.S. stocks. The index accounts for 80% value of the U.S. stock market. Table 2 shows the five stocks with the highest weights in the S&P 500 index in September 2021.

Table 2. Five stocks with the highest weights in the S&P 500 index

Company	Stock symbol	Weight (%)
Apple	AAPL	6.2
Microsoft	MSFT	5.9
Amazon	AMZN	3.9
Facebook	FB	2.4
Alphabet	GOOGL	2.3

Under competitive financial markets, the price of the new asset created by combining underlying assets is precisely the weighted average of the price of each underlying asset. We therefore can synthesize a new augmented asset x_j and its price series by combining underlying stock s_j with weight w_{ij}. Consequently, the price of this augmented asset is $p_{x_j} = \sum_{i=1}^{N} w_{ij} p_{si}$ where p_{si} is the price of stock i, N is the number of underlying stocks/assets. The weight w_i is all positive. Throughout this paper, prices of augmented assets will be constructed in this fashion. Figure 1 shows the price series of a new asset constructed from 0.5 Apple (APPL) stock and 0.5 Microsoft (MSFT) stock. The price of the new asset is the mid point between the prices of APPL and MSFT stocks.
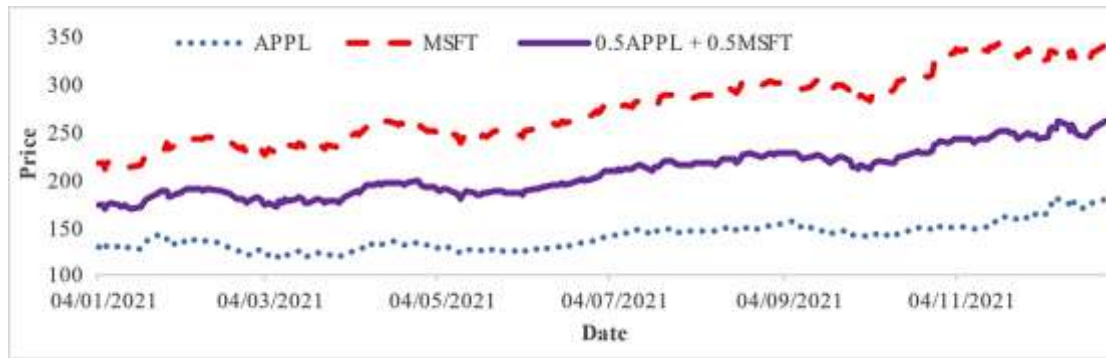
Figure 1. New asset price series

Technically, the data augmentation in (1) is similar to an image data augmentation in [20]. Zhang *et.al.* [20], new images are created by a mixup process: averaging the pixels of two original images. Figure 2 shows a mixup of a dog image and a cat image. This image augmentation brings performance improvement in various datasets. Table 3 shows the differences between the mixup process and our data augmentation. Mixup is applied for image classification problems. Our data augmentation is for stock return regression problems. While mixup has no real-world analog, our data augmentation is based on real-world financial asset creation. The mixup process only applies to two original images. Our augmentation could apply to any number of original stocks greater or equal to 2. Compared with mixup, our augmentation could apply to many more combination of original data points and could potentially generate much larger augmented datasets.

Figure 2. Mixup image of a dog and a cat

Table 3. Mixup and our data augmentation

	Mixup	Our data augmentation
Domain	Images	Times series of stock prices
Real-world foundation	No	Yes
Type of problems	Classification	Regression
Number of data point combined for augmentation	2 images	Any k $\geq$ 2 stock series
Number of all possible combinations of original data points for augmentation (*N* is the number of original data points)	$N(N-1)/2$	$2^N - N - 1$

2.2. Data, feature, and target

The historical adjusted closing prices of each stock in S&P500 from 1 January 2002 to 31 March 2020 are employed for studies. Stocks in S&P500 are extensively used in studies on stock forecasting [21], [22]. The data set will be separated into two parts: 1/1/2002-30/11/2016 and 1/1/2017-31/3/2020. The first and second parts are respectively for training and testing.

The target variable is the 20-day forward returns of each stock. Features employed for predicting future returns are standard technical indicators shown in Table 4. The technical indicators used are rolling volatility, simple moving average normalized with 5-day simple moving average, rolling median to mean, rolling standard deviation to mean, and rolling return. The sizes of rolling windows are 10, 20, 40, 60, 80, 120, and 240 days. Totally, there are 35 features. All the features are generated from the price series of each stock. These features are features commonly used in stock forecasting literature [23]–[25].

Table 4. Feature and target

Variable type	Name	Size of rolling windows (days)
Target	Forward return	20 days
Feature	Rolling volatility	10, 20, 40, 60, 80, 120, 240 days
Feature	Simple Moving Average	10, 20, 40, 60, 80, 120, 240 days
Feature	Rolling median to mean	10, 20, 40, 60, 80, 120, 240 days
Feature	Rolling standard deviation to mean	10, 20, 40, 60, 80, 120, 240 days
Feature	Rolling return	10, 20, 40, 60, 80, 120, 240 days

3.   RESULTS

This section studies the effect of data augmentation on prediction performance measured by root mean square error (RMSE) using augmented data with various sizes. The baseline root mean square errors (RMSEs) with no data augmentation is reported in section 3.1. Then, these RMSEs are compared with those with data augmentation in section 3.2.

3.1. Baseline results without data augmentation

We first show the baseline prediction performance without data augmentation. To focus on data augmentation, we use a standard ML algorithm in all trials. For this purpose, throughout the paper, light gradient boosted machine (LightGBM) with default setting is employed with no hyper-parameter tuning. LightGBM is a well-established ML model developed and maintained by Microsoft. Our data augmentation strategy depends on the number of original stocks (N). To investigate how the data augmentation might be affected by N, we conduct trials with N=10, 20, 30, ..., 100. For each value of N, there are 30 trials. In each trial, N stocks are randomly picked from the S&P 500 stocks. Then the price series of the picked N stocks are used as the primary data source for training and testing in each trial. With 10 values of N, there are totally 10×30 trials.

The mean and standard deviation of RMSEs of the trials for each value of N in test sets is shown in Table 5. As expected, the RMSEs and their standard deviations decrease as the size of training sets and N increases. The decrease in mean and standard deviation (SD) of RMSEs indicates that larger datasets give better prediction performance.

Table 5. Baseline RMSEs

Number of original stocks (N)	Mean of test RMSEs	SD of test RMSEs
10	0.0921	0.0088
20	0.0909	0.0056
30	0.0904	0.0056
40	0.0897	0.0048
50	0.0898	0.0040
60	0.0896	0.0035
70	0.0894	0.0028
80	0.0896	0.0031
90	0.0896	0.0028
100	0.0896	0.0026

3.2. Data augmentation results

We now study learning results with data augmentation. As shown in Figure 3, in each of 10×30 trials with N original stocks, kN additional synthetic assets, and their data are generated and used in training. We experiment with 3 values of k =5, 10, 20. Therefore with data augmentation, there are $10 \times 30 \times 3$ trials. A synthetic price series is created from a random weighted sum of the N original price series. Each weight is randomly distributed according to a standard uniform distribution. After the data augmentation, the LightGBM model is trained using the augmented data. The trained model is then used to predict the forward return of the N original stocks in each trial.

Table 6 reports the average percentage decreases of RMSEs from the models trained with augmented data compared to those without augmented data from trials grouped by the number of original stocks (N) and augmentation ratio (k). All the percentage decreases are positive in Table 6; the prediction accuracy improves in all cases. We apply t-tests to test whether the decreases in RMSEs are different from zero. In all cases, the decreases in RMSEs are highly significantly different from zero. Our data augmentation significantly improves prediction performance for all Ns and ks.

As k increases, the decreases in RMSEs are higher. The larger the additional data from data augmentation, the higher the prediction accuracy. Note that the decreases in RMSEs are larger in training sets

with small N; data augmentation improves learning performance more in training with small datasets. These results show that the performance improved from data augmentation is robust and behaves as expected.
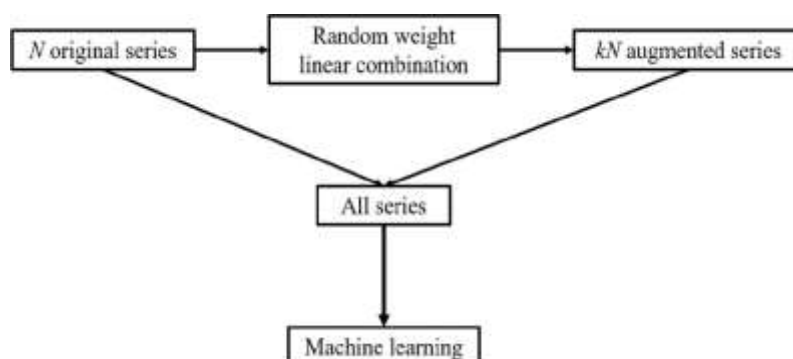

Figure 3. Data augmentation framework

Table 6. Percentage decrease of RMSEs in test set

Number of Stocks (N)	Augmentation ratio (k)		
	5	10	20
10	1.977***	2.527***	2.974***
20	0.041***	0.904***	1.411***
30	0.594***	0.959***	1.250***
40	0.508***	0.748***	1.095***
50	0.484***	0.800***	1.110***
60	0.451***	0.684***	0.967***
70	0.441***	0.652***	0.925***
80	0.475***	0.713***	0.888***
90	0.389***	0.630***	0.900***
100	0.415***	0.649***	0.977***

Significance levels: *p<0.1; **p<0.05; ***p<0.01

3.3. Return correlation and data augmentation performance

In this section, we investigate how the augmentation performance is affected by the characteristics of original stocks. The focus characteristic is the average correlation of original stock returns in each trial. From a financial perspective, correlations of stock returns are one the most important indicators of a portfolio. From a ML perspective, because augmented series are generated by averaging the original stock series, we expect that similarity/difference of the original series could affect the augmentation performance. For example, in an extreme case in which all original series are constant and unrelated, augmenting data by averaging the original series obviously could not improve the learning performance. In Figure 4, the scatter plot of the drop in RMSEs and the average pairwise correlations of the returns of the original N stocks in the training sets of each trial is exhibited. The plots suggest a U-shape relationship of return correlation and the drop in RMSEs. A high drop of RMSEs occurs in the region with high and low return correlations.

To verify the U-shape relationship, linear regressions in which percentage decreases in RMSEs are the dependent variable are estimated. The explanatory variables are return correlations and their squared. The other control variables are the number of original stocks (N) and the augmentation ratio (k). The estimated regression coefficients are shown in Table 7. The first model shows that the augmentation ratio is significantly positive, while the coefficient of N is significantly negative. These two variables have the expected signs. The second model adds return correlation as another explanatory variable. Its coefficient is not significant and shows that the relationship of return correlations and the decreases in RMSEs is not linear. The quadratic term of return correlations is added in the third model. In this model, all coefficients are highly significant with expected signs. This model confirms the non-linear relationship of the decreases in RMSEs and the return correlations of the original stocks; our data augmentation performs better in the region of high and low correlations.
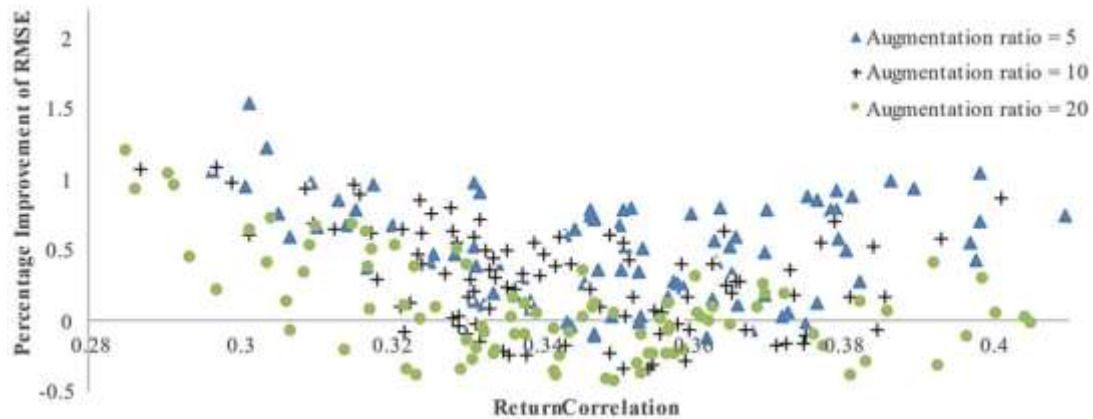
Figure 4. Return correlation and decrease in RMSE

Table 7. Return correlation and decrease in RMSE[a]

Variables	Model		
	I	II	III
# of original stocks (N)	-0.012***	-0.012***	-0.009***
Augmentation ratios (k)	0.041***	0.041***	0.041**
Return correlation		2.001	-112.049***
Return correlation squared			158.650***
Constant	1.120***	0.402	20.657***
R^2	0.177	0.180	0.206
# of observations	900	900	900

[a]Significance levels: *p<0.1; **p<0.05; ***p<0.01

4.   CONCLUSION

This paper proposes a simple and effective data augmentation technique for stock return predictions. New synthetic stocks are generated form linear combinations of original stocks. Unlike existing literature, our technique is intuitive and mimics real asset creation in financial markets. Our data augmentation significantly improves prediction in test sets. The larger the size of augmented data, the larger the improvement. Moreover, regression analysis shows a U-shape relationship between the return correlations of original stocks and the prediction improvement from data augmentation. Our data augmentation works better for groups of original stock with high or low correlation.

REFERENCES

[1]   A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[2]   K. He, X. Zhang, S. Ren, and J. Sun, " Delving deep into rectifiers: surpassing human -level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision* , 2015, pp. 1026–1034, doi: 10.1109/ICCV.2015.123.

[3]   X. Huang, J. Baker, and R. Reddy, "A historical perspective of speech recognition," *Communications of the ACM*, vol. 57, no. 1, pp. 94–103, 2014, doi: 10.1145/2500887.

[4]   R. Grundkiewicz and M. Junczys-Dowmunt, " Near human-level performance in grammatical error correction with hybrid machine translation," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2018, vol. 2, pp. 284–290, doi: 10.18653/v1/n18-2046.

[5]   M. Jaderberg *et al.*, " Human-level performance in 3D multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859–865, 2019, doi: 10.1126/science.aau6249.

[6]   F.-Y. Wang *et al.*, " Where does AlphaGo go: from church-turing thesis to AlphaGo thesis and beyond," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 2, pp. 113–120, Apr. 2016, doi: 10.1109/JAS.2016.7471613.

[7]   C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, Dec. 2019, doi: 10.1186/s40537 -019-0197-0.

[8]   Q. Wen *et al.*, "Time series data augmentation for deep learning: a survey," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Aug. 2021, pp. 4653–4660, doi: 10.24963/ijcai.2021/631.

[9]   H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Data augmentation using synthetic data for time series classification with deep residual networks," *arXiv*, Aug. 2018, doi: 10.48550/arXiv.1808.02455.

[10]   X. Teng, T. Wang, X. Zhang, L. Lan, and Z. Luo, " Enhancing stock price trend prediction via a time-sensitive data augmentation method," *Complexity*, vol. 2020, 2020, doi: 10.1155/2020 /6737951.

[11] S. W. Lee and H. Y. Kim, " Stock market forecasting with super-high dimensional time-series data using ConvLSTM, trend sampling, and specialized data augmentation," *Expert Systems with Applications*, vol. 161, p. 113704, Dec. 2020, doi: 10.1016/j.eswa.2020.113704.

[12] A. Mikolajczyk and M. Grochowski, " Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, May 2018, pp. 117–122, doi: 10.1109/IIPHDW.2018.8388338.

[13] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv*, Dec. 2017, doi: 10.48550/arXiv.1712.04621.

[14] S. Sharifirad, B. Jafarpour, and S. Matwin, " Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs," in *2nd Workshop on Abusive Language Online - Proceedings of the Workshop, co-located with EMNLP 2018*, 2018, pp. 107–114, doi: 10.18653/v1/w18-5114.

[15] J. Wei and K. Zou, " EDA: easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6381–6387, doi: 10.18653/v1/D19-1670.

[16] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, " Audio augmentation for speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Sep. 2015, vol. 2015-January, pp. 3586–3589, doi: 10.21437/interspeech.2015 -711.

[17] D. S. Park *et al.*, " Specaugment: a simple data augmentation method for automatic speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, vol. 2019-Septe, pp. 2613–2617, doi: 10.21437/Interspeech.2019 -2680.

[18] B. K. Iwana and S. Uchida, " An empiricalsurvey of data augmentation for time series classificatio n with neural networks," *PLoS ONE*, vol. 16, no. 7 July, 2021, doi: 10.1371/journal.pone.0254841.

[19] A. Le Guennec, S. Malinowski, and R. Tavenard, "Data augmentation for time series classification using convolutional neural networks," *ECML/PKDD workshop on advanced analytics and learning on temporal data* , 2016. https://halshs.archives-ouvertes.fr/halshs-01357973 (accessed Jun. 21, 2022).

[20] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, " Mixup: beyond empirical risk minimization," Apr. 2018, doi: 10.48550/arXiv.1710.09412.

[21] C. Krauss, X. A. Do, and N. Huck, " Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500," *European Journal of Operational Research*, vol. 259, no. 2, pp. 689–702, 2017, doi: 10.1016/j.ejor.2016.10.031.

[22] J. Uotila, M. Maula, T. Keil, and S. A. Zahra, " Exploration, exploitation, and financial performance: analysis of S&P 500 corporations," *Strategic Management Journal*, vol. 30, no. 2, pp. 221–231, 2009, doi: 10.1002/smj.738.

[23] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, " Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Applied Soft Computing Journal*, vol. 90, 2020, doi: 10.1016/j.asoc.2020.106181.

[24] L. Kumar, A. Pandey, S. Srivastava, and M. Darbari, "A hybrid machine learning system for stock market forecasting," *Journal of International Technology and Information Management*, vol. 20, no. 1, 2011, [Online]. Available: https://scholarworks.lib.csusb.edu/jitim/vol20/iss1/3.

[25] C. J. Neely, D. E. Rapach, J. Tu, and G. Zhou, "Forecasting the equity risk premium: The role of technical indicators," *Management Science*, vol. 60, no. 7, pp. 1772–1791, Jul. 2014, doi: 10.1287/mnsc.2013.1838.