IJETIR
ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Impact Factor: 5.731

Volume 2, Issue 1, September 2020

Deep Learning Model for Financial Time Series Prediction

Mr. B V Pranay Kumar[1], Mr. Avaniketh P[2], Mr. A. Poorna Chandra Reddy[3], Ms. Mallika K. P[4], Mr. Morigadi Aravind[5]

Assistant Professor, Department of Computer Science and Engineering[1,2,3]

Students, Department of Computer Science and Engineering [4,5]

Christu Jyothi Institute of Technology & Science, Jangaon, Telangana, India

Abstract: *Stock market is considered complex, fickle, and dynamic. Undoubtedly, prediction of its price is one of the most challenging tasks in time series forecasting. Traditionally, there are several techniques to effectively predict the next t lag of time series data such as Logistic Regression and Random Forest. With the recent progression in sophisticated machine learning approaches such as deep learning, new algorithms are developed to analyze and forecast time series data. This paper employs Long-Short Term Memory (LSTM) deep learning approach to predict future prices for low, medium, and high risk stocks. To the best of our knowledge, we are proposing an innovating technique to evaluate deep learning and other prediction techniques w.r.t. the stocks' risk factor. The proposed approach is compared with other traditional algorithms over different periods of training data. The results show that our LSTM approach outperforms other traditional approaches for all stock categories over different time periods. Experimental results illustrate that, for low and medium risk stocks, it is better to use LSTM with long time period of training data. However, for high risk stocks, short time period of training data provides more accurate predictions.*

Keywords: Stock market, Machine Learning, Logistic Regression, Random Forest, Deep Learning, Long Short-Term Memory (LSTM), Forecasting, Time Series

I. INTRODUCTION

The stock market is a platform where buyers and sellers trade stocks leading the price either up or down. Stock prices change every minute by market forces. These changes happen because of the law of supply and demand. If more people wantto buy a stock (demand) than sell it (supply), the price moves upward. Conversely, if more people wanted to sell a stock thanbuy it, there would be greater supply than demand, and the price would go down. Investing in the stock market is one of the riskiest decisions made by a person as it may produce in complete loss or massive profit. If the investor wants a high return, he has to be willing to take high-risk and invest in high-risk stocks. If an investor intends to invest in a low-risk stock, he will get a low return. Application of technologieslike machine learning and deep learning aid the investor to grasp the stock price trends and forecasting.

The main objective of this study is to investigate which forecasting methods offer the best predictions with respect to lower prediction errors and higher accuracy of forecasts. Sev- eral studies have been published to predict future stock prices using deep learning algorithms by learning from historical data [15]. Here we proceed in that direction but study a specific method using the Long-Short Term Memory network (LSTM) and compare the results with other traditional methods suchas Logistic Regression and Random Forest.

The experiments use several New York Stock Exchange stocks (NYSE) categorized as low risk, medium risk, andhigh-risk stocks. The outcomes are compared with those for Logistic Regression and Random Forest. The proposed model gives more accurate predictions, as measured by root mean squared error (RMSE) and mean absolute percent error (MAPE). In addition, the proposed approach evinces how much increasing/decreasing training period and increasing the number of epochs for low, medium, and high-risk stock can improve our model.

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 2, Issue 1, September 2020

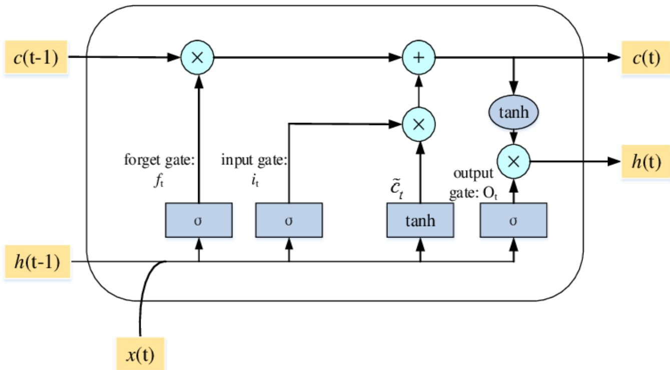


Fig. 1: The internal structure of a Long-Short Term Memory (LSTM) [21].

II. BACKGROUND AND RELATED WORK

Machine learning approaches and more importantly deep learning algorithms have introduced new techniques to pre- diction problems where the relationships between variables are modeled in a deep and layered hierarchy. Machine learning- based techniques such as Logistic Regression (LR) and Ran- dom Forests (RF) have gained a lot of attention in recent yearsin the field of finance.

2.1 Logistic Regression

The standard method for forecasting time series is the statistical linear approach. Logistic regression (LR) model is a type of probabilistic statistical classification model. It is also used to predict a binary response from a binary predictor, usedfor predicting the outcome of a categorical dependent variable(i.e., a class label) based on one or more predictor variables (features). The applications of LR have repeatedly been used in the area of corporate finance, banking, and investments. Logistic regression has the advantage of being less affected than discriminant analysis when the normality of the variable cannot be assumed. Gong and Su [8] explored a new approachbased on Logistic Regression to predict the stock price trendof next month according to the current month. This approach achieved a minimum forecasting accuracy of 83% which is unsatisfactory.

2.2 Random Forests

Random forests algorithm belongs to ensemble learningmethods. The main idea behind ensemble learning methods is that a single classifier is sometimes not sufficient for correct classification of test data and that multiple classifiers can increase model correctness. In which *n* number of trees are constructed while training data, all these trees are made based on different subsets taken from the original training data. Voting is carried out to decide which class to be selected as anoutput class. Random Forests is the most advanced decision tree [10]. Manojlović and Štajduhar [12] used the Random forests algorithm to construct the model used to anticipate five days-ahead and ten days ahead bearings of the CROBEX record. Their outcomes demonstrate that random forests canbe effectively applied for building predictive models for anticipating the course of securities exchange patterns.

2.3 Deep Learning

Deep Learning is a sub-field of machine learning. Deep learning was inspired by artificial neural networks (ANNs). ANNs were inspired by human biological neural networksthat are based on learning by example. Deep learning canbe employed to perform prediction and classification oper- ations based on highly complex training data. Due to the surpassing advances in deep learning, many scientific fields take advantage of its highly accurate performance to build efficient solutions to different types of problems. Deep neural networks (DNNs) showed superior performance in many other areas of applications such as signal processing [9], image classification [11], and speech recognition [22]. Therefore, exploring DNNs techniques in financial time series prediction is a highly recommended model. Many studies have applied different deep learning techniques to time series prediction [15]. A deep recurrent neural network

(RNN) is one of the techniques utilized by many prediction approaches [20].

Deep Learning models are also sensitive to a large numbers of hyper-parameters. In contrast to model parameters which are learned during training, model hyper-parameters are set by the data scientist ahead of training and control implementation aspects of the model. The process of hyper-parameter tuning (also called hyper-parameter optimization) means finding the combination of hyper-parameter values for a deep learning model that performs the best for a specific problem. To tune the hyper-parameters, there are many strategies like random search [18], grid search [17], and COVID-19 propagation model [13]. Deep Learning techniques are proving to be much more accurate as compared to the traditional forecasting techniques. It has been used in text classification, speech recognition, language modeling, and many others. It has also been applied in time series forecasting. Many researchers developed different prediction approaches using LSTM networks. Moghar and Hamiche. [14] have proposed a Long-Short Term Memory model (LSTM) to predict future opening prices for both Google and Nike stocks. They used a four-layer network with different epochs for training data (12 epochs, 25 epochs, 50 epochs, and 100 epochs). It was found that training with fewer data and more epochs can improve the testing result.

Agarwal et al. [5] have presented the application of Long short-term memory (LSTM) network for stock market forecasting using historical data. They used five performance indicators. They took the opening price, closing price, highest price, lowest price, and volume of SBIN stock and achieved an accuracy of 97%, but the sudden fluctuations in the price of many stocks which are derived from abnormal events cannot be factored in the prediction system.

Namini et al. [16] suggested that classical machine learning and deep learning methods for financial metric predictions achieved better results compared to ARIMA models. In ad- dition, they mentioned that the number of training times, known as "epoch" in deep learning, had no effect on the performance of the trained forecast model and it exhibited a truly random behavior. Fischer and Krauss [7] deployed LSTM networks, which are used to carry out the prediction of out-of-sample directional movements for S&P 500 stocks from 1992 to 2015. Yielding daily returns of 0.46 percent, their LSTM networks outperformed memory-free classification (random forest, DNN, etc.).

All these approaches have attempted to predict stocks regardless of the risk factor provided by each of them. In this work, a RNN-based model using the LSTM network is employed to predict stock prices over various periods (5 and 15 years). The study also provides a novel analysis of the effect of changing the training period on the price of stocks with a various risk factors, low, medium, and high risk. Random search is used for hyper-parameters tuning.

III. METHODOLOGY

Long Short-Term Memory (LSTM) is a special kind of Recurrent Neural Network (RNN) capable of learning long-term dependencies. It's also capable of catching data from past stages and using it for future predictions. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods is practically their default behavior, not something they struggle to learn. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell [4]. LSTM is well suited to classify, process, and predict time series given time lags of unknown duration. It trains the model by using back-propagation. The internal structure of an LSTM cell is demonstrated in Figure 1.

This study proposes a long short-term memory model to predict stock price movement. LSTM is an artificial recurrent neural network with additional features to memorize the se- quence of data. The memorization of the earlier trend of the data is possible through some gates along with a memory line incorporated in a typical LSTM. To implement the algorithm, this study used Keras library. The algorithm defines a function called "lstm" that trains and builds the LSTM model. The function takes the training dataset, the number of epochs, i.e., the number of times a given dataset is fitted to the model, the number of neurons, i.e., the number of memory units or blocks, and the number of hidden layers. When compiling a model, a loss function along with an optimization algorithm must be specified. The root mean squared error (RMSE) and Adaptive Moment Es- timation (ADAM) [19] are used as the loss function and the optimization algorithm, respectively.

Random search is applied to setup a grid of hyper-parameter values and select random combinations to train the model and guide the learning process. Random search is great for discovery and getting hyper-parameter combinations that you would not have guessed intuitively. The proposed model uses random search to select the appropriate combination values of number of neurons per layer, learning rate, activation function, batch size, and number of epochs. K-Fold cross validation is applied to avoid the overfitting and estimate the skill of the model on new data (k refers to the number of groups that a given data sample is to be split into).

IV. EXPERIMENTAL SETUP

4.1 Dataset

The daily stock price dataset is collected which consists of closing stock prices of the New York Stock Exchange (NYSE) extracted from yahoo finance. A different number of stocks with different risk levels are used. For low-risk stocks [2] we used A10 Networks (ATEN), Core-Mark Holding Company (CORE), Eli Lilly and Company (LLY), and Chemed Corpo- ration (CHE). For medium-risk [3], we used iShares 20+ Year Treasury Bond ETF (TLT), Vanguard Extended Duration Trea- sury ETF (EDV), iShares Barclays 10-20 Year Treasury Bond Fund (TLH), and Vanguard Long-Term Government Bond ETF (VGLT). For high-risk stocks [1], we used Household robotics company (IRBT), GameStop (GME), Jumia (JMIA), and New Age Beverages (NBEV).

4.2 LSTM model hyper-parameters

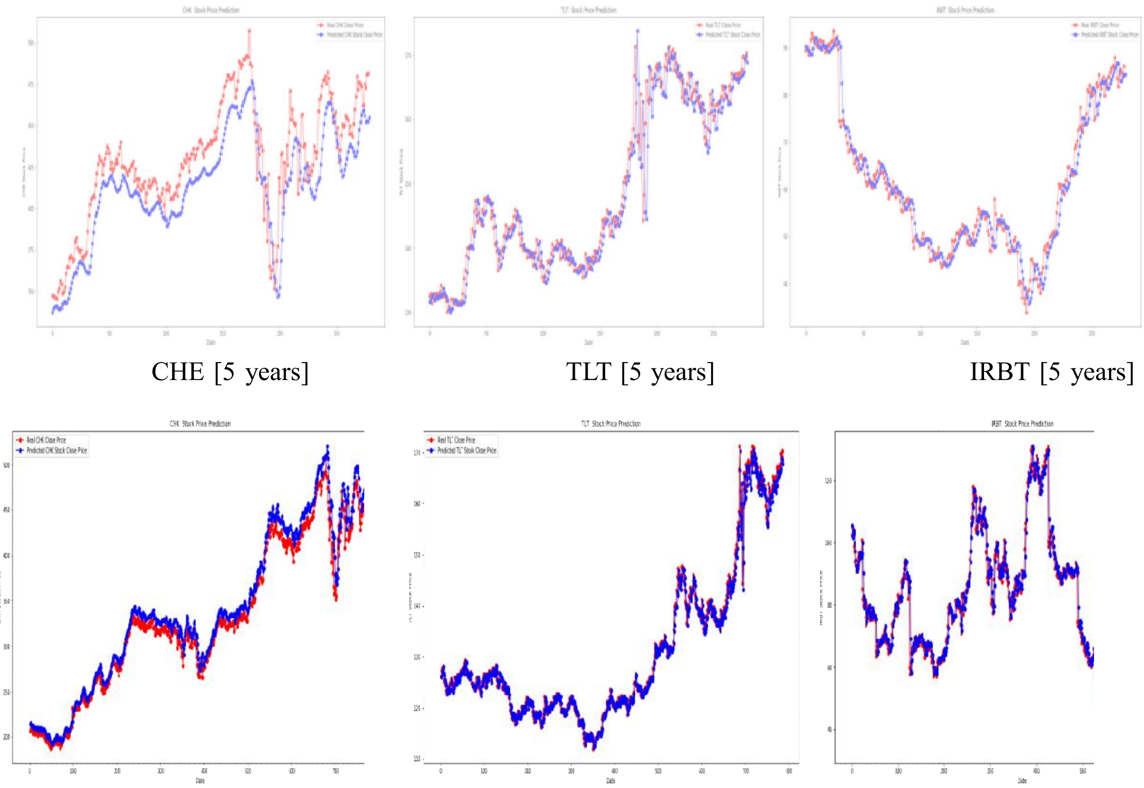
Our LSTM model is composed of 2 layers, one hidden and one output. The hidden layer is with timesteps of 120 days, 50 neurons, batchsize = 32, and tanh is used as activation function. The input weights of the hidden layer are initialized randomly. The output layer contains one neuron to predict the closing price of the next day. Closing price values were normalized in the range of 0–1, before using them as an input for the neural network. This normalization prevents high- value inputs from dominating the final output. We applied different values of training periods (5 years and 15 years). Also, different Epochs for training data are applied (25 epochs, 50 epochs, and 100 epochs). Each financial time series dataset was split into 80-20 subsets where 80% of the data was used for training and the remaining 20% was used for testing. 'Adam' optimizer is applied to minimize the loss function with a learning rate of 0.001. A dense layer connection has been used with a dropout rate of 20 to avoid overfitting.

4.3 Evaluation Methodology

The root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSD represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. RMSE is given in Equation 1,where yi is the observed value and xi is the prediction for the corresponding observed value.

		Deep Learning Long Short Term Memory		Logistic Regression		Random Forest	
		RMSE	MAE	RMSE	MAE	RMSE	MAE
Low Risk	ATEN	**0.220**	**0.158**	1.414	1.0	1.331	0.886
	CORE	**0.9588**	**0.609**	1.4259	1.0166	1.4012	0.9817
	LLY	**0.8718**	**0.697**	1.493	1.115	1.4514	1.0533
MediumRisk	TLT	**1.420**	**0.896**	1.544	0.9172	1.424	0.9833
	TLH	**1.3624**	**0.890**	1.388	0.9637	1.398	0.9779
	VGLT	**1.011**	**0.600**	1.4214	1.0103	1.452	1.0546
HighRisk	GME	**0.5438**	**0.412**	1.468	1.077	1.379	0.9514
	JMIA	**0.6723**	**0.479**	1.4552	1.0588	1.425	1.0158
	NBEV	**0.1369**	**0.100**	1.270	0.806	1.2951	0.8387
Average		**0.799**	**0.537**	1.431	0.996	1.395	0.971

TABLE I: The price prediction error of different stocks data when trained on LSTM, Logistic Regression, and Random Forest,categorized based on risk level.



CHE [5 years] TLT [5 years] IRBT [5 years]

CHE [15 years] TLT [15 years] IRBT [15 years]

Fig. 2: Predicted closing price (Blue line) for IRBT ,GME, and CHE stocks vs. the actual closing value (Red line). First andsecond rows illustrate the training results over 5 and 15 years, respectively.

Mean Absolute Error (MAE) is a popular measure of the prediction accuracy of a forecasting model. It has a very intuitive interpretation in terms of relative error, represented mathematically in equation 2, where y_i is the observed value and x_i is the prediction for the corresponding observed value. It reports the average relative error of the predictions, as a percentage. MAE is useful while evaluating prediction models where only the magnitude of the difference between predicted values and observed values is important to consider [6]. Evaluation using MAE overcomes the large deviation biaspresent in Root RMSE and shows robustness for datasets containing long tails.

V. EXPERIMENTAL RESULTS

5.1 Deep learning (LSTM) vs. Traditional Models

In this section, we discuss the results obtained using our models. Table I shows that LSTM models performed well when compared to traditional models like Logistic Regression

and Random Forest for low, medium, and high-risk stocks. More specifically, the average reduction in error rate obtained by LSTM is 44% lower than Logistic Regression indicating the superiority of LSTM to Logistic Regression. The study shows that LSTM reduced the average error rate by nearly 42% when compared to Random Forest indicating the superiority of LSTM to Random Forest.

5.2 The Impact of Training Periods

The results are reported in Table II and Figure 2. In this study, a series of experiments are performed on low, medium and high-risk stock prices (close price) for different training periods of 5 years from 2 January 2015 to 20 July 2020 and 15 years from 3 January 2005 to 20 July 2020 and captured the error rate. Increasing the number of training periods in medium, and low-risk stocks leads to decreasing the error and improve our testing result however decreasing the number of training periods in high-risk stocks leads to decreasing the error.

International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 2, Issue 1, September 2020

	CHE (Low Risk)		TLT (Medium Risk)		IRBT (High Risk)	
	RMSE	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*
5 years	19.095	16.696	2.5086	1.537	**2.514**	**1.801**
15 years	**13.0579**	**10.72**	**1.420**	**0.896**	3.0574	1.965

TABLE II: The price prediction error of IRBT ,TLT, and CHE stocks data when trained on LSTM over different periods of time (i.e., 5 and 15 years).

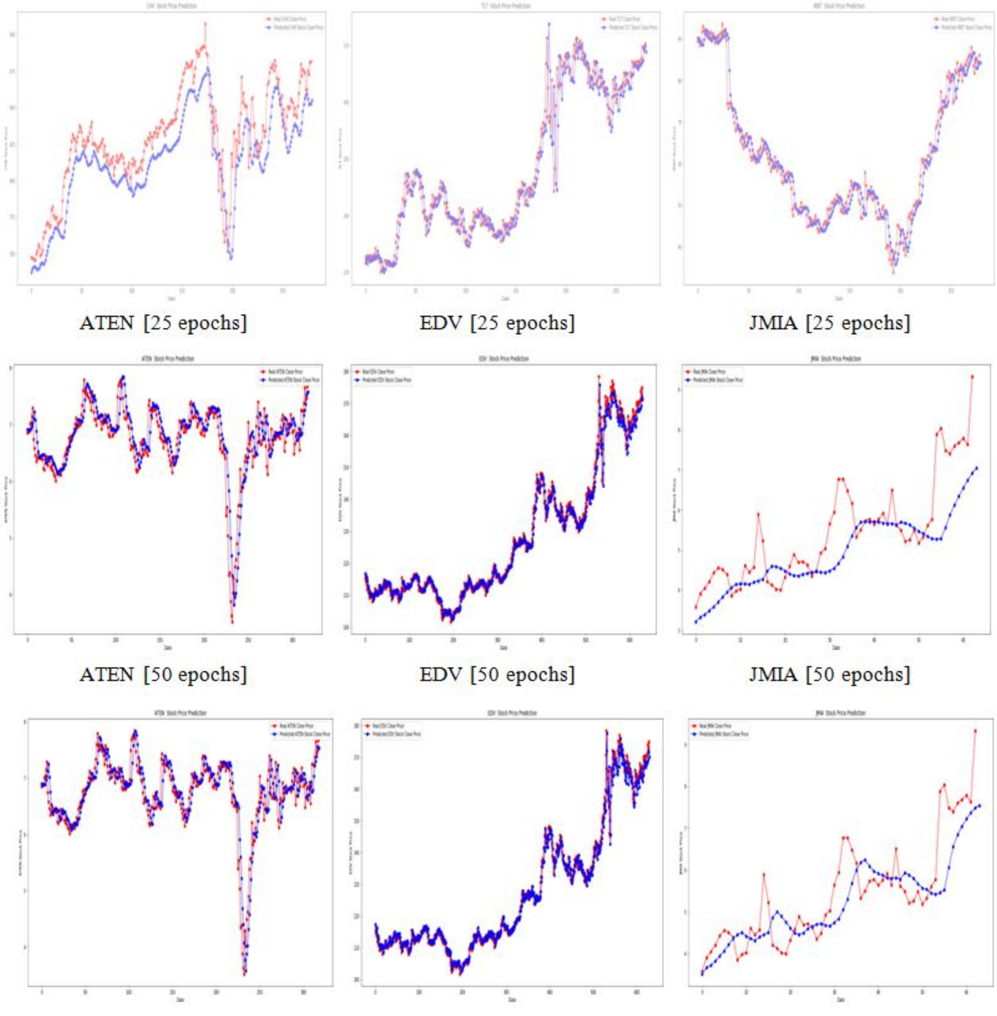


Fig. 3: Predicted closing price (Blue line) for JMIA, EDV, and ATEN stocks vs. the actual closing value (Red line). First,second, and third rows illustrate the number of epochs (i.e., 25, 50, and 100).

	ATEN (Low Risk)		EDV (Medium Risk)		JMIA (High Risk)	
	RMSE	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*
25 epochs	0.3014	0.213	3.3644	2.261	0.9308	0.670
50 epochs	0.251	0.179	2.15	1.379	0.8712	0.622
100 epochs	**0.220**	**0.158**	**2.018**	**1.329**	**0.6723**	**0.479**

TABLE III: The automatic price prediction error of JMIA, EDV and ATEN stocks data when trained on LSTM for 25,50, and 100 epochs

5.3 The impact of the number of Epoch

The study of the influence of the number of training rounds (epochs) on the same data is the focus of this section (Table IIIand Figure 3). A series of experiments and sensitivity analysesare performed in which they controlled the values of the epochand captured the error rate. The epoch values varied between 25, 50, and 100 for each stock. The reason for choosing 100 as the upper-level value for epochs was the practical and feasibility of the experiments. Experimental results show that increasing the number of epochs would decrease the error of the prediction for low, medium, and high-risk stocks.

VI. CONCLUSION

Recently, deep learning has provided state-of-the-art ac- curacy in many fields. With recent progress on developing advanced machine learning-based techniques and in particular deep learning algorithms, these approaches are obtaining popularity among researchers for stock price prediction. The major question is then how accurate and powerful these newly introduced techniques are when compared with traditional methods. This paper provides an LSTM-based approach to predict the prices of stock market prices. We compare our ap- proach to two traditional approaches, Logistic Regression and Random Forests. These three techniques were implemented and applied on a set of financial data over a different periodof time and with various levels of risk (i.e., low, medium,and high-risk stocks). The results indicated that LSTM was superior to both Logistic Regression and Random Forests. More specifically, the LSTM-based algorithm improved the prediction by 44% on average compared to Logistic Regres- sion and 42% compared to Random Forests. Furthermore, the paper reports an improvement when the number of epochs is increased in low, medium, and high-risk stocks regardless of the size of the data. Moreover, Increasing the training periodin low and medium risk stocks leads to a decrease in the error and improve the prediction accuracy. However, decreasing thenumber of training periods in high risk stocks led to a decreasein the percentage of error. These results confirm the nature of high-risk stock as they rapidly change, unexpectedly, over a short period and that make their price prediction less accurate with a long period of training data. In contrast, the low and medium-risk stocks are usually with a fixed or low probabilityof change over a reasonable time interval. Therefore, their prediction accuracy increases with increasing the period of training data.

REFERENCES

[1]. 10 High-Risk stocks that could go either boom or bust. https://finance.yahoo.com/news/10-high-risk-stocks-could- 103442159.html. Accessed: 2020-10-30.

[2]. 5 Low-Risk stocks to buy for epic price swings since 1929. https://finance.yahoo.com/news/5-low-risk-stocks-buy-125012598.html. Accessed: 2020-10-30.

[3]. 8 best long-term investments in october 2020.https://www.bankrate.com/investing/best-long-term-investments/#bondfunds. Accessed: 2020-10-30.

[4]. Colah's blog, understanding lstm networks.http://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed: 2020-10-30.

[5]. Harshit Agarwal, Gaurav Jariwala, and Akshit Shah. Analysis and prediction of stock market trends using deep learning. In Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019), pages 521–531. Springer, 2020.

[6]. Arnaud De Myttenaere, Boris Golden, Be´ne´dicte Le Grand, and Fabrice Rossi. Mean absolute percentage error for regression models. Neuro- computing, 192:38–48, 2016.

[7]. Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2):654–669, 2018.

[8]. Jibing Gong and Shengtao Sun. A new approach of stock price prediction based on logistic regression model. In 2009 International Conference on New Trends in Information and Service Science, pages 1366–1371. IEEE, 2009.

[9]. Reinhold Haeb-Umbach, Shinji Watanabe, Tomohiro Nakatani, Michiel Bacchiani, Bjorn Hoffmeister, Michael L Seltzer, Heiga Zen, and Mehrez Souden. Speech processing for digital home assistants:

Combining signal processing with deep-learning techniques. IEEE Signal processing magazine, 36(6):111–124, 2019.

[10]. Rupesh A Kamble. Short and long term stock trend prediction using decision tree. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), pages 1371–1375. IEEE, 2017.

[11]. Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jo´n Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. IEEE Transactions on Geoscience and Remote Sensing, 57(9):6690–6709, 2019.

[12]. Teo Manojlovic´ and Ivan Sˇtajduhar. Predicting stock market trends using random forests: A sample of the zagreb stock exchange. In 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pages 1189–1193. IEEE, 2015.

[13]. A´ lvarez, Gualberto Asencio-Corte´s, JF Torres, David Gutie´rrez-Avile´s, L Melgar-Garc´ıa, Rube´n Pe´rez-Chaco´n, Cristina Rubio-Escudero, Jose´ C Riquelme, and Alicia Troncoso. Coronavirus optimization algorithm: A bioinspired metaheuristic based on the covid- 19 propagation model. arXiv preprint arXiv:2003.13633, 2020.

[14]. Adil Moghar and Mhamed Hamiche. Stock market prediction using lstm recurrent neural network. Procedia Computer Science, 170:1168–1173, 2020.

[15]. Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing, 90:106181, 2020.

[16]. Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. A comparison of arima and lstm in forecasting time series. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 1394–1401. IEEE, 2018.

[17]. Jose´ F Torres, Antonio Galicia, A Troncoso, and Francisco Mart´ınez- A´ lvarez. A scalable approach based on deep learning for big data time series forecasting. Integrated Computer-Aided Engineering, 25(4):335–348, 2018.

[18]. Jose´ F Torres, David Gutie´rrez-Avile´s, Alicia Troncoso, and Francisco Mart´ınez-A´ lvarez. Random hyper-parameter search-based deep neural network for power consumption forecasting. In International Work-Conference on Artificial Neural Networks, pages 259–269. Springer, 2019.

[19]. Anish Singh Walia. Types of optimization algorithms used in neural networks and ways to optimize gradient descent. Towards Data Science, 2017.

[20]. Pengfei Yu and Xuesong Yan. Stock price prediction based on deep neural networks. Neural Computing and Applications, 32(6):1609–1628, 2020.

[21]. Xiaofeng Yuan, Lin Li, and Yalin Wang. Nonlinear dynamic soft sensor modeling with supervised long short-term memory network. IEEE Transactions on Industrial Informatics, 16(5):3168–3176, 2019.

[22]. Zixing Zhang, Ju¨rgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Bjo¨rn Schuller. Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Transactions on Intelligent Systems and Technology (TIST), 9(5):1–28, 2018.