Advances in Economics, Business and Management Research, volume 128
International Scientific Conference "Far East Con" (ISCFEC 2020)

Spark-Based Big Data Processing: AI to Decide on Opening a SiU9 Contract Position

N I Lomakin[1], P D Kravcheny[1], M C Maramygin[2], A B Shohnech[3], I A Samorodova[4]

[1]Volgograd State Technical University,
Volgograd, Russia
[2]Institute of Finance and Rights of FSBOU VAUD UrGEU-SINH
Yekaterinburg, Russia
[3]Volgograd State Social and Pedagogical University,
Volgograd, Russia
[4]VF FSAOU VO Volga
Volgograd, Russia

E-mail: tel9033176642@yahoo.com

Abstract. This paper presents the theory behind using AI for big data processing. It considers an artificial neural network designed to make decisions whether to open a position on a SiU9 futures contract; to that end, Spark-based big data processing is applied. It is hypothesized and proven herein that the futures closing price can be predicted by the developed AI trained on a dataset containing web-parsed digitized 'news-related' fluctuations as well as time-series Japanese candlesticks of a SiU9 futures contract on a 15-minute timeframe. Research has shown that in the modern world, the ever-larger bulk of stock transactions are done by using trading systems (trading robots). More and more of them use AI. The novelty hereof is that the futures contract price is predicted by an AI trained not only on quantities and Japanese candlestick parameters, but also on the digitized news-related fluctuations collected by Skraper from websites. The dataset generated that was was fed to a perceptron that had a 305-parameter input layer, two hidden layers having (100 parameters and 10 parameters), and an output layer that gave the predicted price. The perceptron was created and run on the Deductor platform, while Skraper was a Python program in the Spark framework. The research team further analyzed how accurately the ANN could predict the price of a financial instrument, the SiU9 futures contract, on a Russian stock exchange over a 15-minute timeframe. The study has effectively produced an ANN-based trading algorithm for opening long and short SiU9 positions in the fixed-term stock market; the algorithm has a good predictive accuracy.

1. Introduction

What makes this research relevant is that AI is extremely important for big data processing, a key factor of accuracy when predicting the price of a stock asset.

The novelty of this research is that it hypothesizes and proves that the futures price can be predicted by the developed AI trained on a dataset containing web-parsed digitized 'news-related' fluctuations as well as time-series Japanese candlesticks of a SiU9 futures contract on a 15-minute timeframe.

Research has shown that in the modern world, the ever-larger bulk of stock transactions are done by using digital trading systems (trading robots), and algorithmic trade is on the rise.

There are many known programs, platforms, and tools for big data analysis. Most popular solutions are frameworks (Hadoop, Spark, Storm), databases (Hive, Impala, Presto, Drill), analytical platforms (RapidMiner, IBM SPSS Modeler, KNIME, Qlik Analytics Platform, STATISTICA Data Miner, Informatica Intelligent Data Platform, World Programming System, Deductor, SAS Enterprise Miner) and other tools (Zookeeper, Flume, IBM Watson Analytics, Dell EMC Analytic Insights Module. This research uses Apache Spark, a framework.

2. Stock-exchange trading robot AI

2.1. Skraper/apache spark big data collection and processing

Big data processing is essential today. In practice, data is collected from the Web by parsing and scraping.

The stock-exchange trading robot uses an AI system based on open-source software: Apache Spark for cluster computing, Apache Cassandra for database storage, Apache MLib for machine learning libraries, Grafana for data analysis and visualization, Apache data streaming and message queue telemetry transport (MQTT) for sensor connectivity [1], see Figure 1.

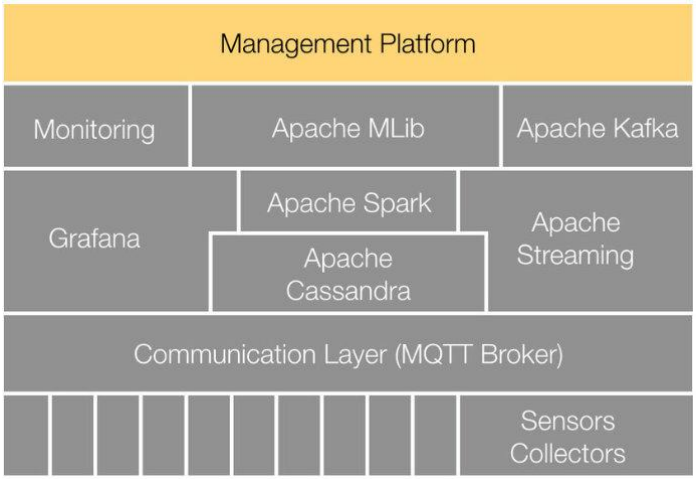

Figure 1. Apache Spark framework [2].

Using the open-source Apache Spark framework enabled the researchers to collect data from news websites over 15-minute timeframe. A list of news URLs was compiled as part of developing Skraper. The idea was to use Spark functionality to count words on web pages.

Skraper was developed to collect news parameters; its algorithm followed a five-block flowchart, see Figure 2.

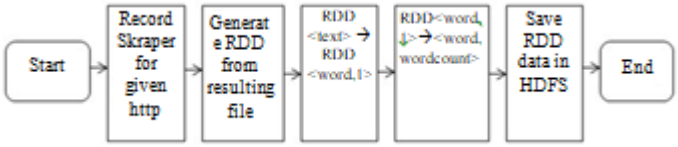

Figure 2. Flowchart of the website scraper.

Scraper was written in Python. Skraper generated text shown in part below, see Figure 3.

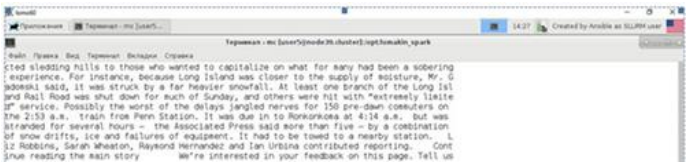

Figure 3. Fragment of Skraper-produced text.

Python WordCount was used to start a resilient distributed dataset for parallelized computing. The obtained unstructured text was transformed and word-counted using RDD <word, 1>-->RDD <word, wordcount>, see Figure 4.

```
('NYTimes.com', 3)
('no', 2)
('longer', 1)
('supports', 1)
('9', 1)
('Please', 4)
('browser.', 1)
('LEARN', 1)
('MORE', 1)
```

Figure 4. Output file format.

Skraper can collect data on word usage patterns observed on news websites over the required timeframe; this data can further be used as intended. For instance, it can be fed to Word2vec, a text-generating app, or to ANN-based ROB advisors, or to convolutional networks for signal recognition.

Textual web news were digitized and converted into the input table as 300-dimensional networks every 15 minutes; each data entry was labeled with SiU9 futures contract data sampled from the QUIK stock-exchange trading platform: the opening price, the maximum price, the minimum price, the closing price, and the traded quantity. Besides, the dataset contained 'predicted' closing prices shifted by one timeframe down. I.e. the actual closing price was the 'predicted' price for the previous timeframe in the training data. Figure 5 below shows some of the input data fed to the AI model.

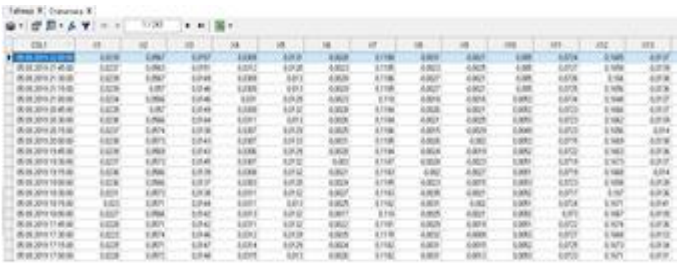

Figure 5. Part of the ANN input.

Theory behind big data processing was important.

2.2. AI-based big data processing: theory

Studies have shown that AI finds ever greater use in big data processing and in other applications, including stock market forecasting.

N.I. Lomakin viewed stock transactions as a factor of growth in the investment activity of real-economy businesses; he also studied the use of stock trading robots in an information society. A research team led by S.P. Sazonov has studied some aspects of using cutting-edge technology in the context of market uncertainty from the standpoint of financial management and its utilization [3]. Competitive advantages are of importance, A.A. Polyanskaya believes [4]. N.I. Lomakin and Ye.V. Loginova analyzed the use of fuzzy algorithms and artificial intelligence in risk management [5]. V.A.

Ekova, O.N. Maksimova, and N.I. Lomakin proposed a systematic approach to enhancing the risk management toolkit [6].

V.A. Vasilyev, A.F. Lyotchikov, and V.E. Lyalin contributed to the stock-exchange financial risk research [7].

AI is used increasingly in all areas of human activity, including stock-exchange trade, predicting the price of a financial asset in a time series, etc. For instance, N. Lomakin, A. Polyanskaya, and Ye. Kharlamova have proposed an ANN to predict profits so as to help regional companies grow sustainably [9].

Trend analysis by S. Sengupta, S. Basak, P. Saikia, S. Paul, V. Tsalavoutis, F. Atiah, V. Ravi, and A. Peters reveals multiple applications for distributed computing systems (Hadoop, Spark) and convolutional neural networks [10]. Distributed computing is becoming ever faster [11,12,13]. Another important trend is that CNN deep learning becomes better [14, 15, 16, 17].

It is essential that multiple authors are studying the issue of CNN speed [18, 18, 20]. Notable are papers by A. Serb and T. Prodromakis, who cover various levels of abstraction for artificial intelligence [21]. Noteworthy are papers by P. Karimi, M.L. Maher, N. Davis, and K. Grace, who studied computational models e.g. deep learning of a computational model for conceptual shifts in the co-design paradigm [22].

Theory behind AI for big data processing is covered in [23]. Of all the programs and tools [24, 25, 26, 27], a subset has been identified for data analysis and machine learning [28, 29, 30, 31].

The authors' coverage of various data mining-based algorithms is of great importance [32]. Note should be taken of such authors as J. Liu, S.J. Gibson, and M. Osadchy, who state that deep learning shows great performance when trained on labeled big data [33]. Noteworthy is N. Udomsak's comparison of the Naive Bayes classifier vs support vector machines in stock-exchange parameter prediction [34], while P. Shiralkar, A. Flammini, F. Menczer, and G.L. Ciampaglia note the need to find threads in knowledge graphs to support fact-checking [35]. The authors hereof have compared the functionality of Hadoop and Spark.

2.3. Stock-exchange trading robot AI

The stock-exchange trading robot AI is designed to help decide whether to open a position on a SiU9 futures contract; to that end, Spark-based big data processing is applied.

Figure 6 below features the time series of SiU9 futures contract prices in USD on a 15-minute timeframe; it also shows the technical analysis graphs.

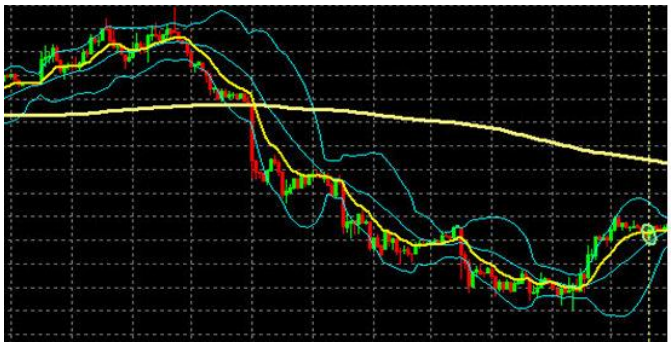

Figure 6. Japanese candlesticks of the SiU9 futures contract; 200-period moving average (bold yellow line), 50-period moving average (thin yellow line), Bollinger bands (turquoise lines) on a 15-minute timeframe.

Studies have shown that the price of a financial instrument cannot be predicted using only the historical parameters of the time series. 'News vibes' are one notable external factor that affects the USD exchange rate.

An ANN was generated and trained to create a trading AI. It used not only the trading quantities or Japanese candlesticks, but also 300 numerical attributes that represented the 'news-related' fluctuations that Skraper collected from the web. The dataset generated that way was fed to a perceptron that had a 305-parameter input layer, two hidden layers (100 parameters and 10 parameters), and an output layer that gave the predicted price.

2.4. Research methods

This research used such methods as the monograph method; calculations; machine collection of data by the author-developed Skraper app for further digitization (300-dimensional vectorization) by Word2Vec in Apache Spark, an open-source framework for distributed processing; Python programming; Deductor-based AI.

3. Results and discussion

3.1. Creating the AI: perceptron to predict SIU9 futures contract price

Skraper-collected news stories from the Internet were fed to Word2Vec, then to the perceptron input as 300-dimensional vectors. These numerical parameters were complemented with Japanese candlestick parameters: opening price, minimum price, maximum price, closing price, and quantities as shown in QUIK. Besides, the dataset contained 'predicted' closing prices shifted by one timeframe down.

Data was split into a training set (95%) and a test set (5%) for the neural network. This effectively created a neural network model: a Deductor-based perceptron, see Figure 7.

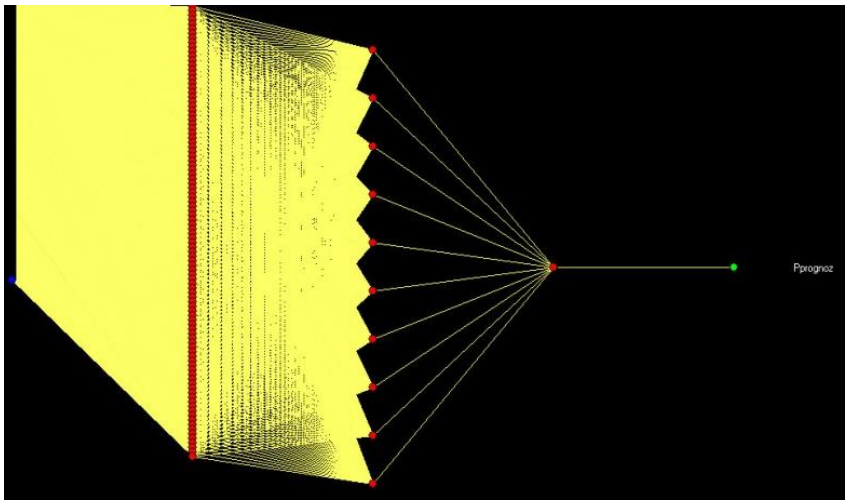

Figure 7. Perceptron flowchart.

3.2. AI performance

The model produced rather good results. Over the analyzed period (5 working days from August 30, 2019, 16:15 to September 5, 2019, 22:00), the yield averaged 32.01% per day. Figure 8 shows the perceptron output.

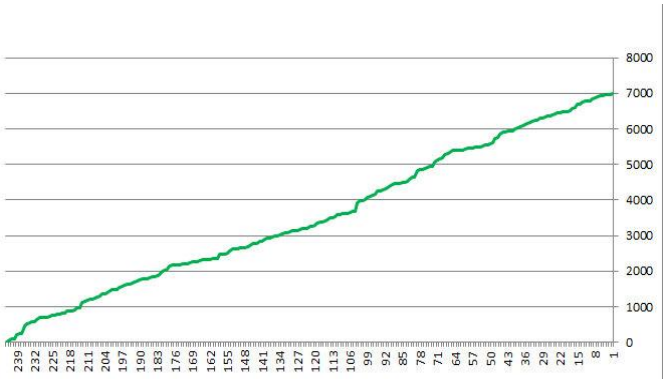

Figure 8. Deposit growth curve for 243 timeframes.

The AI system would open positions at each subsequent timeframe: a short one if the predicted price was below the actual value, or a long one otherwise.

The AI system demonstrated positive return over the course of the 243-timeframe experiment. To trade a single SiU9 futures contract, one needs a minimum sum on their broker account, the so-called collateral, or "the capital" K = 4,059.35 rubles. Testing the algorithm on retrospective data returned a positive margin. The yield amounted to 32.01% a day against the baseline. The accuracy was high, as the predicted SiU9 prices did not deviate from the actual values significantly. Figure 9 below shows the predicted-vs-actual difference.

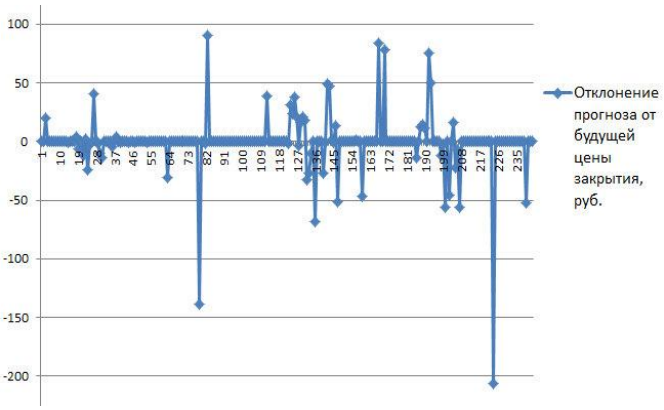

Figure 9. Predicted-vs-actual SiU9 closing prices.

The perceptron had a high yield as the predictive accuracy was high for each 15-minute timeframe, with the mean error (predicted vs actual value) was only -0.48 rubles, see Table 1.

Table 1. Actual and predicted siu9 values.

Date/Time	Price (actual)	Price (predicted)	Difference	Deviation from the actual value (σ)
September 5, 2019, 22:00	66,322	66,483.54	151.54	0.2285
September 5, 2019, 21:45	66,355	66,332.14	-22.86	- 0.0344
September 5, 2019, 21:30	66,360	66,374.74	14.74	0.0222
…	…	….	…	…
Min	6,728.9	6,376.16	-92.92	-0.0138
Max	6,341.9	6,673.79	82.19	0.0129
Middle	6,557.47	6,557.00	-0.48	-0.00007

What-if function was used in Deductor so that the ANN could predict a value.

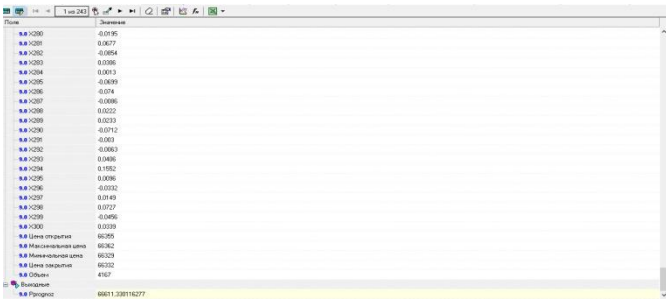

Figure 10. Values predicted using the what-if function.

Table 2 demonstrates the trading robot's performance.

Table 2. AI PERFORMANCE.

Name	Short	Long
Short margin, rubles	3,646	
Long margin, rubles		3,338
Total margin, rubles	6,984	
Collateral (initial capital), rubles	4,059.35	
Yield over 5 business days, times	1.720472	
Number of business days	5	
Number of 15-minute timeframes	243	
Number of 15-minute timeframes per business day	53	
Number of business days	4.584906	
Yield over 1 business day	0.344094	
Commission (2 rubles x number of transactions)	486	
Net margin, rubles	6,498	
Net income per day, %	32.015	

AI finds ever more applications, i.e. predicting Bitcoin prices [36] or risk optimization [37, 38], etc. The conclusions are as follows.

- The paper presents the theory behind the AI-based big data processing. It considers an artificial neural network designed to make decisions whether to open a position on a SiU9 futures contract; to that end, Spark-based big data processing is applied.
- It is proven herein that the futures closing price can be predicted by the developed AI trained on a dataset containing web-parsed digitized 'news-related' fluctuations as well as time-series Japanese candlesticks of a SiU9 futures contract on a 15-minute timeframe.
- The yield amounted to 32.01% a day against the baseline. The accuracy is high, as the predicted SiU9 prices did not deviate from the actual values significantly.

4. Conclusion

Studies have shown that AI finds ever greater use in big data processing and in other applications, including stock market forecasting. The authors' coverage of various data mining-based algorithms is of great importance.

The AI system would open positions at each subsequent timeframe: a short one if the predicted price was below the actual value, or a long one otherwise.

It is hypothesized and proven herein that the futures closing price can be predicted by the developed AI trained on a dataset containing web-parsed digitized 'news-related' fluctuations as well as time-series Japanese candlesticks of a SiU9 futures contract on a 15-minute timeframe.

5. Acknowledgments
Research was supported by the RFBR *Grant for the Development of Innovation and Investment Policy as a Concept of Strategic Economic Security of Agricultural Companies in the Context of Technological Transformation* Today, 19-010-00985 A.

References

[1] ResearchGate URL: https://www.researchgate.net
[2] Apache Spark ™ is a unified analytical engine for processing large amounts of data URL: https://spark.apache.org/
[3] Sazonov S P, Harlamova E E, Yezangina I A, Gorshkova N, Kovajenkov M A, Polyanskaya E A Theory and Methodology of the Financial Management of the Regional Supporting *University Journal of Advanced Research in Law* 8 **1** pp 211-219
[4] Sazonov S P, Harlamova E E, Gorshkova N V, Polyanskaya E A 2016 Competitive advantages of the regional support university and its role in the regional development strategy *Science and Society* **3** vol 1 pp 180-189
[5] Lomakin N I, Loginov E V 2014 Risk management of the EEP financial system based on Fuzzy-algorimes and artificial intelligence systems In the collection: Management of strategic potential of regions of Russia: methodology, theory, practice collection of reports of the All-Russian scientific and practical conference Responsible editor: A V Kopylov pp 196-197
[6] Ekova V A, Maksimova O N, Lomakin N I 2016 Improvement of tools for managing sustainable development of the region *Russian entrepreneurship* T 17 **23** pp 3347-3364
[7] Vasilyev V A, Pilots A F, Lyalin V E 2006 Mathematical models of assessment and risk management of economic entities *Audit and financial analysis* **4** pp 200-237
[8] Felmer G, Sheed A 2008 Introduction to stochastic finance Discrete time/Per with English (M.: MCMNO) 496 p
[9] Lomakin N I 2018 Sustainable development of regional enterprises based on the neural network profit forecasting model Proceedings of the International Scientific Conference "Competitive, Sustainable and Secure Development of the Regional Economy: Response to Global Challenges" (CSSDRE 2018) (Volgograd, Russia, 18-20 April, 2018) ed. by E G Russkova Higher School of Economics, Department of World Economy, Volgograd State University, Institute of Economics and Finance *Publisher: Atlantis Press* pp 113-116 URL : https://www.atlantis-press.com/proceedings/cssdre-18/publishing.
[10] Sengupta S, Basak S, Saikia P, Paul S, Tsalavoutis V, Atiah F, Ravi V and Peters A 2019 A Review of Deep Learning with Special Emphasis on Architectures, Applications and Recent Trends
[11] Konstantinidis K and Ramamoorthy A 2019 Resolvable Designs for Speeding up Distributed Computing
[12] Qian J, Sengupta S and Hansen L K 2019 Active Learning Solution on Distributed Edge Computing
[13] Severinson A, Graell Amat A, Rosnes E 2018 Block-Diagonal and LT Codes for Distributed Computing With Straggling Servers
[14] Nwankpa C, Ijomah W, Gachagan A and Marshall S 2018 Activation Functions: Comparison of trends in Practice and Research for Deep Learning
[15] Vandal T, Kodra E, Dy J, Ganguly S, Ramakrishna N and Ganguly A R 2018 Quantifying Uncertainty in Discrete-Continuous and Skewed Data with Bayesian Deep Learning
[16] Shen J, Liu J, Chen Y, Li H 2019 Towards Efficient and Secure Delivery of Data for Deep Learning with Privacy-Preserving

[17] Zhang H, Wang H, Chen X, Wang Y and Jin Y 2018 OMNIRank: Risk Quantification for P2P Platforms with Deep Learning
[18] Konstantinidis K and Ramamoorthy A 2019 Resolvable Designs for Speeding up Distributed Computing
[19] Eggensperger K, Lindauer M, Hutter F 2017 Neural Networks for Predicting Algorithm Runtime Distributions
[20] Wang J, Liu J, Pu J, Yang Q, Miao Z, Gao Y Song 2019 An alarm prediction framework for financial IT system using hybrid machine learning methods
[21] Serb A and Prodromakis T 2019 A system of different layers of abstraction for artificial intelligence
[22] Karimi P, Maher M L, Davis N and Grace K 2019 Deep Learning in a Computational Model for Conceptual Shifts in a Co-Creative Design System, 6/24/2019cs.HC, cs. LG, stat. ML
[23] Li D, Ouyang B, Wu D, Wang Y 2019 Artificial intelligence empowered multi-AGVs in manufacturing systems
[24] Elkano M, Sanz J, Barrenechea E, Bustince H and Galar M 2019 CFM-BD: a distributed rule induction algorithm for building Compact Fuzzy Models in Big Data classification problems
[25] Dai J, Wang Y, Qiu X, Ding D, Zhang Y, Wang Y, Jia X, Zhang C, Wan Y, Li Z, Wang J, Huang S, Wu Z, Wang Y, Yang Y, She B, Shi D, Lu Q, Huang K, Song G 2018 BigDL: A Distributed Deep Learning Framework for Big Data
[26] Makkie M, Huang H, Zhao Y, Vasilakos A V and Liu T 2017 Fast and Scalable Distributed Deep Convolutional Autoencoder for fMRI Big Data Analytics
[27] Hossein A and Rahnama A 2016 Distributed Real-Time Sentiment Analysis for Big Data Social Streams
[28] Boukouvalas Z, Elton D C, Chung P W and Fuge M D 2018 Independent Vector Analysis for Data Fusion Prior to Molecular Property Prediction with Machine Learning
[29] Gupta A, Thakur H, Shrivastava R, Kumar P, Nag S 2017 A Big Data Analysis Framework Using Apache Spark and Deep Learning
[30] Kochura Y, Stirenko S, Alienin O, Novotarskiy M, Gordienko Y 2017 Performance Analysis of Open Source Machine Learning Frameworks for Various Parameters in Single-Threaded and Multi-Threaded Modes
[31] Kochura Y, Stirenko S, Rojbi A, Alienin O, Novotarskiy M, Gordienko Y 2017 Comparative Analysis of Open Source Frameworks for Machine Learning with Use Case in Single-Threaded and Multi-Threaded Modes
[32] Kumar P, Kumar N V, Durg S, Chauhan S 2012 A Benchmark to Select Data Mining Based Classification Algorithms For Business Intelligence And Decision Support Systems
[33] Liu J, Gibson S J and Osadchy M 2018 Learning to Support: Exploiting Structure Information in Support Sets for One-Shot Learning
[34] Udomsak N 2015 How do the naive Bayes classifier and the Support Vector Machine compare in their ability to forecast the Stock Exchange of Thailand?
[35] Shiralkar P, Flammini A, Menczer F and Ciampaglia G L 2015 Finding Streams in Knowledge Graphs to Support Fact Checking
[36] Lomakin N I, Poletavkina T A, Salygina I I, Sukhorukov N N, Lukyanov G I, Maximova O N, Gorbunova A V, Maly N A, Burdyugova O M, Golodova O A, Ivanova A V 2018 Use of a neuronet for forecasting of the price of Bitcoin Science Krasnoyarya T 7 **1-2** pp 81-89
[37] Ustun B and Rudin C 2019 Learning Optimized Risk Scores
[38] Solovev D B, Gorkavyy M A 2019 Current Transformers: Transfer Functions, Frequency Response, and Static Measurement Error *2019 International Science and Technology Conference "EastConf", International Conference on.* [Online]. Available: https://doi.org/10.1109/EastConf.2019.8725351