A forecasting of stock trading price using time series information based on big data

Soo-Tai Nam[1], Chan-Yong Jin[2], Seong-Yoon Shin[3]
[1]Institute of General Education, Pusan National University, Republic of Korea
[2]Division of Information and Electronic Commerce, Wonkwang University, Republic of Korea
[3]School of Computer Information and Communication Engineering, Kunsan National University, Republic of Korea

Article Info

Article history:

Received Jul 3, 2020
Revised Sep 27, 2020
Accepted Oct 8, 2020

Keywords:

Big data
Data mining
Euclidean distance
Forecasting
Stock trading
Time series

ABSTRACT

Big data is a large set of structured or unstructured data that can collect, store, manage, and analyze data with existing database management tools. And it means the technique of extracting value from these data and interpreting the results. Big data has three characteristics: The size of existing data and other data (volume), the speed of data generation (velocity), and the variety of information forms (variety). The time series data are obtained by collecting and recording the data generated in accordance with the flow of time. If the analysis of these time series data, found the characteristics of the data implies that feature helps to understand and analyze time series data. The concept of distance is the simplest and the most obvious in dealing with the similarities between objects. The commonly used and widely known method for measuring distance is the Euclidean distance. This study is the result of analyzing the similarity of stock price flow using 793,800 closing prices of 1,323 companies in Korea. Visual studio and Excel presented calculate the Euclidean distance using an analysis tool. We selected "000100" as a target domestic company and prepared for big data analysis. As a result of the analysis, the shortest Euclidean distance is the code "143860" company, and the calculated value is "11.147". Therefore, based on the results of the analysis, the limitations of the study and theoretical implications are suggested.

Corresponding Author:

Seong-Yoon Shin
School of Computer Information and Communication Engineering
Kunsan National University
558 University-Ro, Kunsan-City, 54150, Republic of Korea
Email: s3397220@kunsan.ac.kr

1. INTRODUCTION

Recently, due to the proliferation of mobile and the introduction of web services, not only online structured data, but also unstructured data is rapidly increasing, and it is used in various ways in various fields [1]. In the case of big data, the annual average growth rate of 23.1% is expected from 2014 to 2019 in the global market, and the annual average growth rate of 26.4% from 2014 to 2018 is expected in the domestic market [1, 2].

In particular, the emergence of social media in the field of big data has been an opportunity for the rapid spread and accumulation of unstructured data accumulated from individuals and organizations regardless of time and place. In fact, about 70% of recently generated digital data is generated in various social media where users generate data, including e-mail [2-4]. A good advantage of this more accurately predict a diversified contemporary society, and can provide personalized information to individual. In order to extract meaningful information from a large number of unstructured data generated in social media, interest in big data technology is increasing in various fields, and continuous discussions are being made on how to effectively manage and analyze big data [5, 6].

Generally, big data refer to a large amount of large data beyond the range that can be stored, managed, and analyzed by existing database software. However, it is difficult to simply define big data on a volume. Big data describe large scale data that include not only structured data, but also unstructured data types such as text, image, video, and voice. ig data generated in various environments has a large data size compared to general data, and the data creation speed is very fast [7, 8]. It is said that big data has three characteristics: volume of data, velocity of data creation, and variety of information types [9, 10]. In conclusion, the three aspects are generally called "V", and recently "3V" is also defined as "4V" including the value of the fourth aspect, big data. And, scholars called the oil of the 21st century big data. Efficient refining of crude oil can produce high added value raw materials like gasoline.

Therefore, it can be profitable to extract valuable information from a large number of data. Big data can be used to solve various problems in the general enterprise. Analysis of big data will help you to operate and manage your company. Big data technology into existing data management and analysis system indicates the technique used to gain insight from the huge extent of the data difficult to handle. Google is the most notable company with big data.

Today, the emergence of big data brings a variety of changes to the way of life in human. The development of computer and information communication technology (ICT) has made it possible to analyze big data. In addition, the importance of big data as a core resource and tool in various fields such as industrial, public, medical, and science, especially in developed countries, is emerging [11]. However, one of the problems that continues to be mentioned with the positive future prospects of big data is related to invasion of personal privacy and protection of personal information.

In the big data era, digital data such as location information, search patterns, and access records generated and generated through various smart devices is generated. In addition, even in the case of data created and released at the will of the person, the possibility of the infringement of personal information continues to increase as such information is used or abused in an unintended direction [12-14]. With such problems, research on big data analysis and research on big data security has been actively conducted in certain fields.

The time series data are obtained by collecting and recording the data generated in accordance with the flow of time. Such time series data occurs not only in science, but also in various fields such as medicine, economic, and medical care. If the analysis of these time series data, found the characteristics of the data implies that feature helps to understand and analyze time series data. In particular, the problem of finding meaningful features of the time series data collected in the past and using them to predict future data changes has long been of interest to many researchers. The concept of distance is the simplest and the most obvious in dealing with the similarities between objects. The Euclidean distance is the most widely used methods of measuring the distance between objects, Minkowski distance, Manhattan distance, Mahalanobis distance, Chebyshev distance and Hamming distance.

2. RESEARCH METHOD

Big data is a set of data that goes beyond the ability of common database management tools to capture, store, manage, and analyze. Recently, due to the spread of mobile and the introduction of web services, the amount of online data has been rapidly increasing and is being used in various fields. In particular, the advent of social media in the field of big data has triggered a rapid increase in the amount of unstructured data that has been accumulated. In order to extract meaningful information from these unstructured data, there is increasing interest in big data technology in various fields [3, 4]. A good advantage of this more accurately predict a diversified contemporary society, and can provide personalized information to individual. And, scholars called the oil of the 21st century big data. Efficient refining of crude oil can produce high added value raw materials like gasoline. Therefore, it can be profitable to extract valuable information from a large number of data. Big data can be used to solve various problems in the general enterprise. Analysis of big data will help you to operate and manage your company.

Analysis and forecasting of the stock market have long been recognized as a very important research project, not only in the economic field, but also in mathematics, statistics, and computation. Recently, with the development of financial engineering, research on the prediction and the use of stock prices through scientific methods has been greatly activated. The stock price prediction algorithm is classified into three types: mathematical prediction, statistical predictive, and artificial intelligence prediction. Recently, in order to compensate for the weaknesses of financial engineering systems, patterns are extracted from SNS or news and applied to stock price prediction [15, 16].

First, mathematical prediction is a technique that predicts the future value quantified based on a mathematical model to determine whether to invest, such as building a portfolio or trading. The Black Sholes Model, published by Fisher Black and Myron Shoals in 1973, became the basis for all options trading, and various techniques have since emerged. Representatively, the filtration method (percolation method) that studies how the price moves on a trading order with a limited transaction price range. The wavelet transform is used to analyze the movement of time series data and use it to predict the association between data and future motion.

There is a moving average analysis that divides the arithmetic average of stock prices within a certain period and expresses them as the average stock price. There is a Monte Carlo simulation method that statistically obtains a stochastic distribution of the results to be obtained by generating a large number of random numbers. Statistical forecasting is an approach to predict the future based on historical stock market data. The AI-based stock price prediction method, which began in the late 1980s, finds optimized parameters applicable to predictive models. SVM, ANN, and GA are widely used in classification and regression analysis. It is widely used to find optimal patterns or weighting variables of predictive models using neural networks or genetic algorithms.

Prediction using SNS or news is a method of extracting meaningful features in a document through text mining processing after collecting text data. Using this, it is classified whether the news is good or bad for the stock price and then attempts to predict the simulation investment and price fluctuation using the classification result. Bollen [16] predicted the rise and fall of the dow jones indices (DJIA) by measuring six emotion modes (calm, alert, sure, vital, kind, happy) detected by Twitter. Schumaker [15] proposed AZFinText, a machink learning system that derives stock price predictors from the news, and conducted experiments that simulated trading.

The concept of distance is the simplest and most obvious in dealing with the similarity between an object and an object in a specific coordinate or space. K-nearest neighbor algorithm is used for classification learning and is a very simple and efficient nonparametric method proposed by Hart in 1968. It is a very intuitive method of finding the k-nearest individuals in the training dataset for a single entity based on the similarity between the samples and assigning the highest frequency group within the k-sets. There are many ways to measure similarity within k-nearest neighbors. Euclidean distance, Minkowski distance, Manhattan distance, Mahalanobis distance, Chebyshev distance and Hamming distance are the most widely used and widely known Euclidean distances [17, 18]. In general, the One dimension is a vertical line. The Two dimensions represent the coordinate plane and the Three dimensions represent the space plane. The most commonly used and widely known of these is the Euclidean distance. Therefore, based on this methodology, Figure 1 shows three equations for measuring the distance between entities [19-22]. We want to analyze the similarity between entities using raw data collected using the following equation.

◈ One Dimension (vertical line)

$A(x_1), B(x_2), \qquad d = \sqrt{(x_2 - x_2)^2}$

◈ Two Dimension (coordinate plane)

$A(x_1, y_1), B(x_2, y_2), \qquad d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

◈ Three Dimension (space)

$A(x_1, y_1, z_1), B(x_2, y_2, z_2), \qquad d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$

Figure 1. Equation of measuring distance for each dimension

3. RESULTS AND DISCUSSION

It is common for most entrepreneurs, economists, investors, general asset owners, and general equity investors to be concerned with the price of a stock rather than the value of the original company. Although this researcher is not a stock expert with extensive knowledge, he is one of those who are interested in investing in stocks [23-25]. The purpose of this study is to analyze the similarity of stock transaction prices between companies using 793,800 data of 1,323 companies listed on the stock market in Korea. In addition, the data used in this study are based on specific daily closing prices. No environmental variable intervention other than a specific daily closing price was used in this analysis. Based on the research methodology mentioned above, the basic data used in this study are posted as shown in Table 1. The A1 (column) used for data analysis has meaning in the code of a specific company in Korea. The row should be interpreted as the meaning of the daily closing price of a particular company in Korea. In order to conduct accurate research, it’s necessary to perform preprocessing of data before analyzing the data.

Table 1. Raw data for stock closing prices of companies in Korea

Firms	Day-1	Day-2	Day-3	Day-4	Day-5	Day-6	Day-7	Day-8	Day-9	Day-10	Day…
000100	230500	226000	227500	226000	224500	220000	220000	220000	221500	221500	224500
000105	168000	165500	167000	167500	169500	169500	169000	170000	170000	170000	170000
000120	166000	169000	170000	166500	167000	165500	167000	167500	168500	168500	170000
000140	11400	11500	11650	11550	11550	11700	11750	11550	11650	11550	11500
000145	10100	10100	10350	10300	10550	10600	10600	10250	10600	10650	10550
000150	128000	128000	129000	130000	126000	127000	125500	127000	125500	127500	128500
000155	78400	78100	78300	78700	78000	77800	77500	77700	78400	77900	76500
000157	77400	77800	78000	78100	77600	77600	77400	77600	77000	76600	75900
000180	2960	2915	2915	2900	2880	2880	2870	2880	2890	2855	2895
000210	85400	83800	85000	86200	86800	86300	85300	85600	86100	87000	86900
000215	33950	33600	33600	34150	34150	34700	34300	34600	34600	35100	35300
000220	12200	11800	12150	12400	11900	11550	11650	11650	11750	11800	12000
000225	7320	7240	7190	7100	7100	7050	7110	7140	7120	7180	7340
000227	26050	27000	28000	26700	26000	25800	25450	24800	25350	26700	28750
000230	15100	15150	15200	15100	15250	14800	14750	14600	14700	14800	15050
000240	21050	21300	21100	21300	21150	21050	20950	20650	20600	20850	20750
000270	35550	35450	36750	35750	35900	35350	34800	35200	35500	35700	35350
000300	892	896	900	903	901	896	907	906	913	918	925
000320	17850	17300	17900	18000	18000	17800	17800	17800	17350	16750	17000
000325	21950	20900	20500	20550	20550	20300	19800	20350	20000	20000	20000
000327	19600	19600	19600	19600	19100	20250	20500	20700	21800	22150	22500
000370	9230	9690	9680	9570	9550	9810	9710	9550	9690	9500	9990
000390	8650	8730	8860	8950	8940	8990	9000	8990	8930	8710	8700
000400	4345	4475	4290	3970	3995	3875	3830	3825	3800	3620	3635
000430	3990	4105	4110	4115	4130	4100	4090	4120	4135	4170	4190
000480	89600	89500	89300	89100	88200	88600	88500	88800	88900	88800	88700
000490	7290	7260	7250	7340	7340	7390	7230	7200	7350	7380	7380
000500	23600	23600	23350	23000	23550	23600	23350	24100	23600	23100	23900
000520	8110	8030	8130	8080	7920	8010	7960	7950	8000	8010	7950
000540	6970	7130	6880	6690	6650	6730	6480	6480	6500	6400	6490
000545	6830	6920	6940	6700	6670	6800	6800	6800	6830	6840	6850
000547	24600	25200	25000	25050	24800	25050	24700	24750	24400	24300	23550
000590	82000	83600	83400	84100	84500	85200	85200	85100	84500	84600	84700
000640	126500	126500	126500	126500	126500	126500	126500	126500	126500	126500	126500
000650	96400	96500	96200	95600	95700	95800	95000	95000	95000	95200	95300
000660	68700	68600	68500	67600	67900	68400	68100	68000	68200	67200	67000
000670	122000	123500	126700	127300	127600	127600	126600	128300	131900	129700	131600
	0	0	0	0	0	0	0	0	0	0	0
000680	3145	3135	3125	3150	3150	3150	3165	3190	3180	3195	3270
000700	7670	7760	7760	7670	7800	7550	7800	7910	7930	7890	7820
000720	40750	40650	41700	42700	42550	42450	41950	42150	41650	41250	41800
000725	54100	55100	55100	54900	55700	56200	56500	56600	58200	57700	57100
000760	18850	18700	18750	19050	18900	18900	18850	19000	18800	18850	19400
000810	270000	276500	281500	286500	285000	292500	288500	283000	285000	284000	286500
000815	186000	187000	187000	188000	190000	192500	193500	191000	192000	194000	195500
000850	60700	60500	60200	60300	60000	59600	59900	60200	57800	59200	57700
000860	37700	37750	37950	38500	38650	38700	38800	38650	39000	38800	39500
000880	48650	49150	49200	49350	49300	49200	51400	52000	51800	50600	51700
000885	26500	26700	26700	26300	26600	26800	27950	28250	28250	27950	28050

First, normalization work on daily prices should be performed. If you do not do the preceding work, you may not find meaningful results [25-27]. After the normalization data preprocessing process, we performed full scale data analysis. We decided to measure the similarity of the stock price flow between companies by the Euclidean distance method. Based on the analysis results, we have drawn a graph of the stock price of the two companies to help readers understand. The following Figure 2 is the stock price graph of “000100” company. And, the following Figure 3 is the stock price graph of “143860” company. It can be seen that the stock price flow graphs of the two companies below are similar when viewed from a visual perspective.

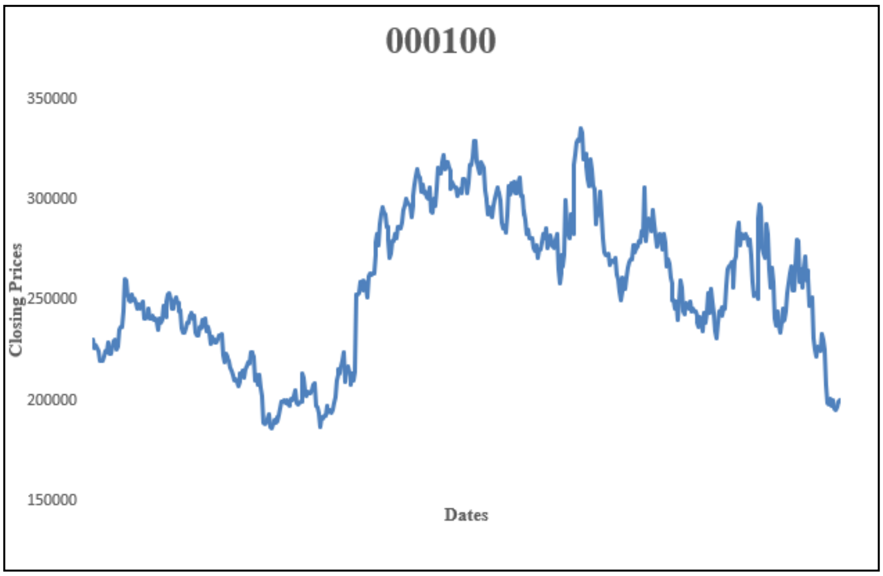

Figure 2. Stock price graph of company “000100”

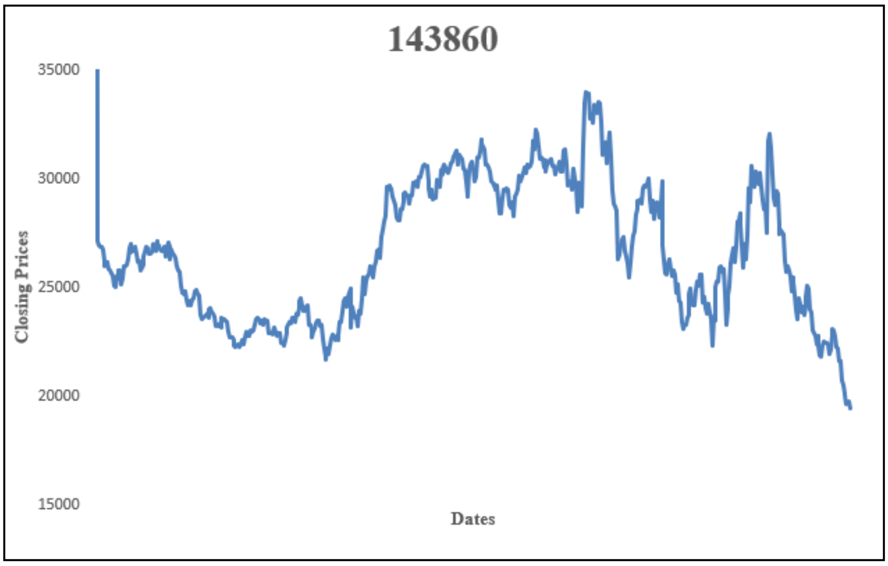

Figure 3. Stock price graph of company “143860”

4. CONCLUSION

Finding meaning through based on big data analysis is the ultimate goal of all researchers. Things you never thought of in the past are now possible. Today, in the big data era, it is possible to forecast the future through data. Based on previous studies, similarity was analyzed through stock trading prices using a simple and clear Euclidean distance. This study is the result of analyzing the similarity of stock price flow using 793,800 closing prices of 1,323 companies in Korea. As a result, Euclidean distance is a method of classifying similar companies using the price flow of stocks between companies. We calculated the Euclidean distance after coding using visual studio as the most convenient and smart big data analysis tool. First, we selected “000100” as a target domestic company and prepared for big data analysis. Next, Euclidean distances for 1,323 companies were calculated based on the reference company using visual studio. As a result of the analysis, the shortest Euclidean distance is the code “143860” company, and the calculated value is “11.147”. The meaning of Euclidean distance is interpreted as showing similar stock price flows in the past. It can’t be said that the two companies with these results show the same share price trend in the future. However, it can be said that the flow of stock prices seems to be similar. Finally, Figures 2 and 3 show the company graphs of similar stock price flows with Euclidean distance calculations. If we compare the graphs of the two companies closely, we can say that they are not the same, but similar. We were not trying to find the same company. However, it means that both companies provide understanding through the price flow of stocks in the past. It must be very difficult to predict the future from past records. However, sometimes the same phenomenon occurs.

REFERENCES

[1] S. M. Rue, "BigData Effects on Artificial Intelligence," *Journal of Korean Institute of Information Technology,* vol. 14, no. 1, pp. 29-34, 2016.

[2] Y. J. Jang and S. K. Cho, "A Comparative Analysis of Data Gathering and Sampling Methods for Social Data," *Social Science Studies,* vol. 25, no. 2, pp. 3-25, 2014.

[3] S. T. Nam, et al., "A Reconstruction of Classification for Iris Species Using Euclidean Distance Based on a Machine Learning," *Journal of the Korea Institute of Information and Communication Engineering,* vol. 24, no. 2, pp. 225-230, 2020.

[4] K. J. Santosh, "Performance evaluation of Map-reduce jar pig hive and spark with machine learning using big data," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 10, no. 4, pp. 3811-3818, 2020.

[5] S. H. Yun, et al., "The Method of Digital Copyright Authentication for Contents of Collective Intelligence," *Journal of the Korea Convergence Society,* vol. 6, no. 6, pp. 185-193, 2015.

[6] J. H. Kim, et al., "A Scheme of Social Engineering Attacks and Counter measures Using Big Data based Conversion Voice Phishing," *Journal of the Korea Convergence Society,* vol. 6, no. 1, pp. 85-91, 2015.

[7] J. Manyika, et al., "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute,* 2011.

[8] G. H. Han and S. H. Jin, "Introduction to Big Data and the Case Study of Its Applications," *Journal of The Korean Data Analysis Society,* vol. 16, no. 3, pp. 1337-1351, 2014.

[9] P. Carter, "Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO," *White pa-per, IDC sponsored by SAS,* pp. 1-16, 2011.

[10] A. McAfee, "Big data, The management revolution," *Harvard Bus Rev,* vol. 90, no. 10, pp. 61-67, 2012.

[11] K. S. Noh and J. Y. Lee, "A Study on Analysis of the Differences for Perception of Big Data in Era of Convergence," *Journal of Digital Convergence,* vol. 13, no. 10, pp. 305-312, 2015.

[12] C. W. Park, et al., "An Empirical Research on Information Privacy Risks and Policy Model in the Big data Era," *The Journal of Society for e-Business Studies,* vol. 21, no. 1, pp. 131-145, 2016.

[13] H. S. Lee, et al., "Personal Information Overload and User Resistance in the Big Data Age," *The Journal of Intelligence and Information Systems,* vol. 19, no. 1, pp.125-139, 2013.

[14] B. C. Kim, "Big Data Security Technology and Response Study," *Journal of Digital Convergence,* vol. 11, no. 10, pp. 445-451, 2013.

[15] R. P. Schumaker, and H. Chen. "A discrete stock price prediction engine based on financial news," *COMPUTER-IEEE Computer Society,* vol. 43, no. 2, pp. 51-56, Jan. 2010.

[16] J. Bollen, et al., "Twitter mood predicts the stock market," *Journal of Computational Science,* vol. 2, no. 1, pp. 1-8, 2011.

[17] Y. M. Chun and S. S. Jeong, "A Comparison of Euclidean Distance and Grey Relational Grade," *Journal of the Korean Data Analysis Society,* vol. 9, no. 2 pp. 687-702, Apr. 2007.

[18] Y. Liu, et al., "Model Design for Reduce OTP Reauthorization Based on Euclidean Distance," *Journal of Knowledge Information Technology and Systems,* vol. 12, no. 5, pp. 737-745, Oct. 2017.

[19] S. Y. Shin, and H. C. Lee, "Realistic Enhancement of 3D Expressions for Building Expressions with Hologram," *Journal of the Korea Institute of Information & Communication Engineering,* vol. 23, no. 09, pp. 1104-1109, 2019.

[20] H. M. Lee, and S. Y. Shin, "Design of The Wearable Device considering ICT-based Silver-care," *Journal of the Korea Institute of Information & Communication Engineering,* vol. 22, no. 10, pp. 1347-1354, 2018.

[21] S. P. Kim, and J. M. Kim, "A Study on Open Source Software Business Model based on Value," *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology,* vol. 7, no. 2, pp. 237-244, 2017.

[22] G. Wang and S. Y. Shin, "An Improved Text Classification Method for Sentiment Classification," *Journal of information and communication convergence engineering,* vol. 17, no. 1, pp. 41-48, 2019.

[23] X. Tan and S. Y. Shin, "Differential Evolution with Multi-strategies based Soft Island Model," *Journal of Information and Communication Convergence Engineering,* vol. 17, no. 4, pp. 261-266, 2019.

[24] D. J. Park and W. S. Kim, "Improvement of the Parallel Importation Logistics Process Using Big Data," *Journal of Information and Communication Convergence Engineering,* vol. 17, no. 4, pp. 267-273, 2019.

[25] S. R. Salkuti, "A survey of big data and machine learning," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 10, no. 1, pp. 575-580, Feb. 2020.

[26] P. H. Huynh, et al., "Enhancing Gene Expression Classification of Support Vector Machines with Generative Adversarial Networks," *Journal of information and communication convergence engineering,* vol. 17, no. 1, pp. 14-20, 2019.

[27] X. Yang, et al., "Fault Diagnosis Management Model using Machine Learning," *Journal of information and communication convergence engineering,* vol. 17, no. 2, pp. 128-134, 2019.

BIOGRAPHIES OF AUTHORS

Soo-Tai Nam is is a Lecturer in the Institute of General Education, Pusan National University, South Korea. Thus, in the Wonkwang University and Kunsan National University, South Korea. He completed his Ph. D. degree in Department of Information Management from Wonkwang University, South Korea, and has 10 years of teaching and researching experience. His interests include MIS, E-Business, Technology Management, Big-Data, Internet of Things.

Chan-Yong Jin is a Professor in the Department of E-Commerce & Inforrmation Management at Wonkwang University, South Korea. He completed his Ph. D. degree in Management Information System from Seonam University, South Korea, and has 30 years of teaching and researching experience. His interests include Meta-Analysis, Big data and E-Commerce.

Seong-Yoon Shin is a Professor in the School of of Computer Information Engineering of Kunsan National University, South Korea. He completed his Ph. D. degree in Computer Science from Kunsan National University, South Korea, and has 30 years of teaching and researching experience. His research interests include Image Processing, Computer Vision, and Multimedia.