This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779181, IEEE Access

An Empirical Study on Modeling and Prediction of Bitcoin Prices with Bayesian Neural Networks Based on Blockchain Information

Huisu Jang and Jaewook Lee.

***Abstract*—Bitcoin has recently attracted considerable attention in the fields of economics, cryptography, and computer science due to its inherent nature of combining encryption technology and monetary units. This study reveals the effect of Bayesian neural networks (BNNs) by analyzing the time series of Bitcoin process. We also select the most relevant features from Blockchain information that is deeply involved in Bitcoin's supply and demand and use them to train models to improve the predictive performance of the latest Bitcoin pricing process. We conduct the empirical study that compares the Bayesian neural network with other linear and non-linear benchmark models on modeling and predicting the Bitcoin process. Our empirical studies show that BNN performs well in predicting Bitcoin price time series and explaining the high volatility of the recent Bitcoin price.**

***Index Terms*—Bitcoin, Blockchain, Bayesian neural network, Time-series analysis, Predictive model**

I. Introduction

BITCOIN is a successful cipher currency introduced into the financial market based on its unique protocol and Nakamoto's systematic structural specification [1]. Unlike existing fiat currencies with central banks, Bitcoin aims to achieve complete decentralization. Participants in the Bitcoin market build trust relationships through the formation of Blockchain based on cryptography techniques using hash functions. Inherent characteristics of Bitcoin derived from Blockchain technologies have led to diverse research interests not only in the field of economics but also in cryptography and machine learning.

Numerous studies have been conducted recently on modeling the time series of Bitcoin prices as a new market variable with specific technical rules. Generalized Autoregressive Conditional Heteroskedasticity (GARCH) volatility analysis is performed to explore the time series of Bitcoin price [2], [3]. Various studies on statistical or economical properties and characterizations of Bitcoin prices refer to its capabilities as a financial asset; these research focus on statistical properties [4], [5], inefficiency of Bitcoin according to efficient market hypothesis [6], [7], hedging capability [8], [9], speculative bubbles in Bitcoin [10], the relationship between Bitcoin and search information, such as Google Trends and Wikipedia [11], and wavelet analysis of Bitcoin [12].

Relatively few studies have thus far been conducted on estimation or prediction of Bitcoin prices. [13] evaluates Bitcoin price formation based on a linear model by considering related information that is categorized into several factors of market forces, attractiveness for investors, and global macro-financial factors. They assume that the first and second factors mentioned above significantly influence Bitcoin prices but with variation over time. The same researchers limit the number of regressors to facilitate linear model analysis. [14] predicts the Bitcoin pricing process using machine learning techniques, such as recurrent neural networks (RNNs) and long short-term memory (LSTM), and compare results with those obtained using autoregressive integrated moving average (ARIMA) models. A machine trained only with Bitcoin price index and transformed prices exhibits poor predictive performance. [15] compares the accuracy of predicting Bitcoin price through binomial logistic regression, support vector machine, and random forest.

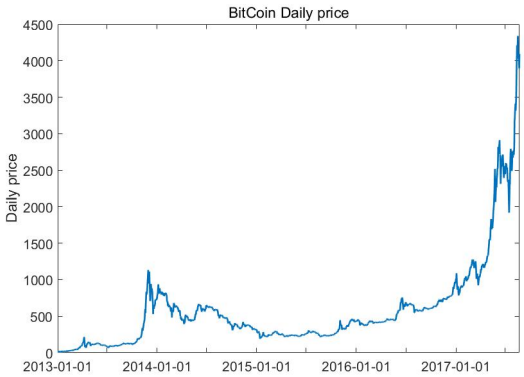


Fig. 1. Bitcoin daily price(USD), from Sep-11 2011 to Aug-22 2017

There are few practical and systematic empirical studies on the analysis of the time series of Bitcoin. In this study, we conduct practical analysis on modeling and predicting of the Bitcoin process by employing a Bayesian neural network (BNN), which can naturally deal with increasing number of relevant features in the evaluation. A BNN includes a regularization term into the objective function to prevent the overfitting problem that can be crucial to our frame-

Huisu Jang and Jaewook Lee are with Department of Industrial Engineering, Seoul National University. e-mail: *gmltn7798@snu.ac.kr* (Huisu Jang), and *jaewook@snu.ac.kr* (Jaewook Lee)

Please address all correspondences to Dr. Jaewook Lee, Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea.

Manuscript received September 20, 2017.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779181, IEEE Access

work. When the machine considers a lot of input variables, a trained machine can be complex and suffer from the overfitting problem. BNN models showed their effect to the financial derivative securities analysis [16]. Formation of Blockchain, a core technology of Bitcoin, distinguishes Bitcoin from other fiat currencies and is directly related to Bitcoin's supply and demand. To the best of our knowledge, in addition to macroeconomic variables, direct use of Blockchain information, such as hash rate, difficulties, and block generation rate, has not been investigated to describe the process of Bitcoin price. To fill this gap, the current study systematically evaluates and characterizes the process of Bitcoin price by modeling and predicting Bitcoin prices using Blockchain information and macroeconomic factors. We also try to account for the remarkable recent fluctuation, which is shown in Figure 1 and has not been considered in previous studies.

The rest of this article is structured as follows: Section II describes Bitcoin and Blockchain technique, which is a distinctive feature of Bitcoin not included in other general currencies. Section III briefly reviews the BNNs employed to model the process of Bitcoin prices. Section IV presents the experimental design and data specifications. Section V outlines empirical results. Section VI concludes the paper.

II. Bitcoin and Blockchain

A. Economics of Bitcoin

Barro's model [17] provides a simple Bitcoin pricing model under perfect market conditions as in [13]. In this model, Bitcoin is assumed to possess currency value and is exchangeable with traditional currencies, which are under central bank control and can be used for purchasing goods and services. The total Bitcoin supply, S_B, is represented by

$$S_B = P_B B \tag{1}$$

where P_B denotes the exchange rate between Bitcoin and dollar (i.e. dollar per unit of Bitcoin), and B is the total capacity of Bitcoins in circulation.

The total Bitcoin demand depends on the general price level of goods or services, P; the economy size of Bitcoin, E; and the velocity of Bitcoin, V, which is the frequency at which a unit of Bitcoin is used for purchasing goods or services. The total demand of Bitcoin,D_B, is described as followed by:

$$D_B = \frac{PE}{V} \tag{2}$$

The market equilibrium with the perfect market assumption is acquired when the supply and the demand of Bitcoin is the same amount. The equilibrium is therefore achieved at

$$P_B = \frac{PE}{VB} \tag{3}$$

This equilibrium equation implies that in the perfect market, the Bitcoin price in dollars is affected proportionally by the general price level of goods or services multiplied by the economy size of Bitcoin, and inversely by the velocity of Bitcoin multiplied by the capacity of the Bitcoin market. The general price level of goods or services, P, can be determined indirectly from the global macroeconomic indexes in actual markets. The exchange rate between several fiat currencies and Bitcoin price describes the relationship between actual markets and Bitcoin market. The main difference between the Bitcoin market and general currency markets originates from the fact that the Bitcoin is a "virtual currency based on Blockchain technologies". Therefore, economic size, E; the velocity, V; and the capacity of the Bitcoin market, B, are closely related with several measurable market variables extracted from the Blockchain platform and, which will be reviewed in the next subsection.

B. Blockchain

Decentralization is the value pursued by all cryptocurrencies as opposed to general fiat currencies being valued by central banks. Decentralization can be specified by the following goals: (i) Who will maintain and manage the transaction ledger? (ii) Who will have the right to validate transactions? (iii) Who will create new Bitcoins? The blockchain is the only available technology that can simultaneously achieve these three goals. Generation of blocks in the Blockchain, which is directly involved in the creation and trading of Bitcoins, directly influence the supply and demand of Bitcoins. Combination of Blockchain technologies and the Bitcoin market is a real-world example of a combination of high-level cryptography and market economies.

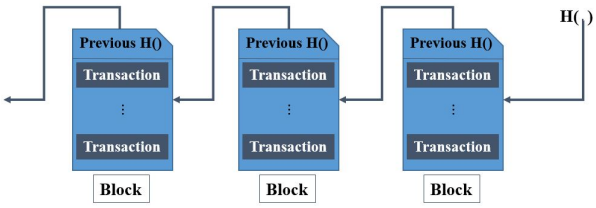


Fig. 2. The formation of the Blockchain

We then describe in detail how the Blockchain can achieve the abovementioned goals in Bitcoin environment [18]. A participant in a Bitcoin network acts as a part of a network system by providing hardware resources of their own computer, which is called a "distributed system". All issuance and transaction of money are conducted through P2P networks. All trading history is recorded in the Blockchain and shared by the network, and all past transaction history is verified by all network participants. The unit called "block", which includes recent transactions and a hash value from the previous "block", creates irreversible data by a hash function, and is pointed out from the next block. Figure 2 shows the general structure of Blockchain. It takes more than a certain amount of time to generate the block to make impossible to forge all or part of the Blockchain. This algorithm is called proof of work (PoW), and the difficulty is automatically set to ensure that the problem can be solved within approximately 10 minutes. PoW also provides incentives to motivate participants to

maintain the value of Bitcoin by paying Bitcoin for the participant who created the block.

PoW agreement algorithm comes with several inherent risks. First, the validity of the block can be intervened when the majority of total participants is occupied by a group with a specific purpose called 51% problem. Second, when the Blockchain is forked, a considerable amount of time is consumed to form the agreed Blockchain until the longest chain is selected after generation of several blocks. This condition causes a transaction delay because the transaction cannot be completed during that time. Lastly, there may be the capacity limit of the Blockchain or the performance limit of each node. Safety of the current Blockchain can be monitored by observing measurable variables in the Blockchain from https://blockchain.info/.

Considering that supply and demand of Bitcoin are affected directly or indirectly by measurable variables involved in the formation of a Blockchain, the current study evaluates several variables related to Blockchain formation as features of the Bitcoin pricing process. Section IV describes in detail the variables exploited in empirical experiments.

III. Time series modeling

For time series analysis, nonlinear methods, such as kernel regression model, exponential autoregressive models, artificial neural network (ANN), BNN, and support vector regression, have attracted research interest and exhibited improved predictive performance for various time series data [16], [19]–[26].

[22] demonstrated that Nikkei 225 index future options in 1995 were better predicted by neural networks using the back-propagation algorithm than the traditional Black-Scholes models. [16] showed that generalization for pricing and hedging derivatives can be improved by the Bayesian regularization techniques and verified empirically for S&P 500 index daily call options from January 1988 to December 1993. [23] reported that support vector regression (SVR) improved the forecast accuracy for the daily currency market data of AUD/USD, EUR/USD, USD/JPN, and GBP/USD options from January to July in 2009. [24] presented support vector regression methods optimized by chaotic firefly algorithm outperforms several methods of SVR for NASDAQ quotes, Intel (from 9/12/2007 to 11/11/2010), National Bank shares (from 6/27/2008 to 8/29/2011) and Microsoft (from 9/12/2007 to 11/11/2011) daily closed stock prices. [26] tuned the parameters of multi-output support vector regression using firefly algorithm and compared the proposed SVR methods with other existing methods for forecasting the market indexes, S&P 500, Nikkei 225, and FTSE 100 indexes.

[25] showed that the least squares support vector machines has better prediction performance for the time series of electrical energy consumption of Turkey compared to the traditional regression models and artificial neural networks. [21] proposed the time series prediction methods combining backpropagation neural networks and least squares support vector machines for the time series prediction for depth-averaged current velocities of underwater gliders.

Unlike other widely studied time series researches, there are few related papers analyzing the Bitcoin processes in terms of prediction performance. In this work, we have employed Bayesian neural networks since the predicted model with a large number of input variables need to be regularized for the weights. We have compared a prediction performance of BNN methods with linear regression methods and SVRs, which are representative prediction methods using various input variables. In this section, we describe a Bayesian neural networks model employed for our experiments.

A. Bayesian neural networks

Bayesian neural networks (BNN) is a transformed Multilayer perceptron (MLP) which is a general term for ANNs in the fields of machine learning. The networks have been successful in many application such as image recognition, pattern recognition, natural language processing, and financial time series [27]. It becomes known that much effective to represent the complex time series than the conventional linear models, i.e. autoregressive and moving average, etc. The structure of a BNN is constructed with a number of processing units classified into three categories: an input layer, an output layer, and one or more hidden layers.

Specifically, neural networks containing more than one hidden layers can solve the exclusive OR (XOR) problem, which cannot be solved by a single layer perceptron [28]. Different from a single layer perceptron, which can only be linearly separated, they solve XOR problems by introducing backpropagation algorithms and hidden layers. The hidden layer mapping the original data to a new space transforms data that cannot be linearly separated into linearly separable data.

Weights of a BNN must be learned between the input-hidden layer and hidden-output layer. Backpropagation refers to the process in which weights of hidden layers are adjusted by the error of hidden layers propagated by the error of the output layer. An optimization method called delta rule is used to minimize the difference between a target value and output value when deriving backpropagation algorithm. In general, BNNs minimize the sum of the following errors, E_B, using backpropagation algorithm and delta rule.

$$E_B = \frac{\alpha}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} (t_{nk} - o_{nk})^2 + \frac{\beta}{2} {}_B^T {}_B \tag{4}$$

where E_B is the sum of the errors, N is the number of the training variables, K is the size of the output layer, t_{nk} is the k-th variable of the n-th target vector, o_{nk} is the k-th output variable of the n-th training vector, α and β are the hyper-parameter, and ${}_B$ is the weights vector of the Bayesian neural network.

A BNN is a non-linear version of ridge regression, which is largely based on the Bayesian theory for neural networks. Unlike conventional neural networks that maximize

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779181, IEEE Access

marginal likelihood, BNN is a machine maximizing the value of posterior through an application of the Bayes' theory. The elements added to the error term cause the machine to learn by selecting a weight with high importance even when the number of total weights is reduced rather than distributed to a large number of weights.

B. Resampling methods

In this section, we discuss two representative resampling methods: *cross-validation*, and *bootstrap*. We identify advantages and disadvantages of each method and select the appropriate method for the empirical analysis of this study.

A *bootstrap method* is one of the sampling techniques that new data set is sampled from the original data set with the replacement. A typical bootstrap works as follows [29]:

1 We have the original data set$D$ with the number of N.
2 Below following step is repeated B times for particular large number to produce B different bootstrap data set, $Z_1, Z_2, \cdots, Z_B$
   - Data set Z_i with the size N is generated by sampling from the original data set D with the replacement.
3 The machine is trained from each bootstrap data set.
4 Accuracy of the machine is calculated by averaging each bootstrap data set.

$$Accuracy = \frac{1}{B}\sum_{j=1}^{B}\frac{1}{N}\sum_{i=1}^{N}(1 - Loss(\hat{y}_i^j, y_i)) \tag{5}$$

where y_i is an i-th true training output data, $\hat{y}_i^j$ is an i-th estimated output from the bootstrap data Z_j, and $Loss(\cdot,\cdot)$ is a loss function.

A *cross-validation* randomly divides the original data set into K equal-sized parts without the replacement. We fit the machine learning model to the $K-1$ parts leaving out particular set k and acquire a prediction error for the left-out k part. Total prediction accuracy is combined after the procedure is repeated for each part to leave [30], [31]. A general procedure is as follows:

1 We divide the original data set into K partial equal-sized data set,$C_1, C_2, \cdots, C_K$, without the replacement. n_k is the number of each partial set defined by n/K.
2 We can compute the total accuracy:

$$accuracy_K = \sum_{k=1}^{K}\frac{n_k}{N}\frac{1}{n_k}\sum_{i=1}^{n_k}(1 - Loss(\hat{y}_i^k, y_i)) \tag{6}$$

where N is the total number of the original data set, others have same definition with in the bootstrap description.
3 The estimated standard deviation of the cross-validation:

$$\hat{SE}(CV_K) = \sqrt{\frac{\sum_{k=1}^{K}(Err_k - \bar{Err}_k)^2}{N-1}} \tag{7}$$

where Err_k is the k-th loss, $\sum_{i=1}^{n_k} Loss(\hat{y}_i^k, y_i)$.

Bootstrap is adequate to validate a predictive model performance, to use an ensemble method, and to estimate of bias and variance of the trained model. Bootstrap creating the cloned multiple samples with the replacement is not originally developed for model validation. It can give more biased results. Therefore, we employ the cross-validation technique to our model validation. Cross-validation can create high-variance problems when data size is small. Our data size is sufficient to overcome the problem. We employ the 10-fold cross-validation methods generally used for model validations.

IV. Blockchain data description

This section describes Blockchain data and macroeconomic variables used in our empirical analysis and their summary statistics.

A. Data specification

Figure 1 shows the time series of Bitcoin price obtained from https://bitcoincharts.com/markets/, where the value of 1-Bitcoin, which was about $ 5 in September 2011, approximates $ 4,000 in August 2017. During this period, market volatility with enormous price changes in Bitcoin becomes exceptional compared with that in traditional currency markets. It is evident that standard economic theories are insufficient to account for the impressive price development and volatility of Bitcoin [11]. Bitcoin markets do not possess purchasing power nor interest rate parity. In particular, Bitcoin is an actual implementation of decentralization issued under the consent of participants and not the central bank. This fact suggests that the need for completely new determinants of Bitcoin price: *the Blockchain information* that includes relevant features as main determinants for pricing Bitcoin. Blockchain data used for empirical analysis can be collected from https://blockchain.info/. Table I presents the Blockchain data and macroeconomic variables to be used in predicting the evolution of Bitcoin prices.

TABLE I
Data for the empirical study

Data category	Data
Response var.	prices or log prices of Bitcoin(USD), vol. or log vol. of Bitcoin(USD)
Blockchain information	Trading vol.(USD,CNY), avg. block size, transactions/block, median confirm. time, hash rate, difficulty, cost % of trans., miners' rev., confirmed trans., total num. of uniq. Bitcoin
Macro economic development	S& P500, Eurostoxx, DOW30, NASDAQ, Crude oil, SSE, Gold, VIX, Nikkei225, FTSE100
Global currency ratio(/USD)	GBP, JPY, CHF, CNY, EUR

Several blockchain variables are considered as follow:

- **Average block size (MB)**: the size of a block verified by all participants.

- **Transactions per block**: average number of transactions per block.
- **Median confirmation time**: the median time for each transaction to be accepted into a mined block and recorded to the ledger.
- **Hash rate**: estimated number of Tera (trillion) hashes per a second all miners (market participants to solve a hash problem for making a block) is performing.
- **Difficulty**: next difficulty =(previous difficulty $*2016*$ 10 minutes)/(time to mine last 2016 blocks)
- **Cost % of a transaction**: miners' revenue as the percentage of the transaction volume.
- **Miners revenue**: Total value of coin-base block rewards and transaction fees paid to miners.
- **Confirmed transaction**: the number of confirmed the validity of transactions per day.
- **Total number of a unique Bitcoin**: market capitalization of Bitcoin.

By employing ordinary least square (OLS) estimation, [32] demonstrates that the Dow Jones index, the euro-dollar exchange rate, and WTI oil price influence the value of Bitcoin price in the long run. We also consider several variables such as S& P500, Eurostoxx, DOW30, NASDAQ, Crude oil, SSE, Gold, VIX, Nikkei225, and FTSE100, which associated with global macroeconomic development.

Given that Bitcoin is related to traditional currency markets in addition to the cryptocurrency market itself based on digital cryptography, we take into account the exchange rates between global monetary markets; exchange rates are basic factors in the analysis of traditional currency markets. We specifically use *exchange rates between major fiat currencies* (GBP, JPY, CHF, CNY, EUR) and the dollar because these rates are most likely to affect the Bitcoin price.

In summary, we cover the daily data from Sep 11, 2011, to Aug 22, 2017 in the empirical analysis by employing both the traditional determinants of currency markets, such as global macro-economic development and the features endowed from the cryptocurrency. This experiment, which has not been performed in previous studies, primarily aims to discover the main features that can explain the recent highly volatile Bitcoin process.

B. Summary data statistics

Table II shows summary statistics of response variables, Blockchain-related variables, global macroeconomic indexes, and international exchange rates used in empirical analysis from September 13, 2011, to July 21, 2017. Several notable points are considered in the empirical analysis. As shown in Table II, response variables and Blockchain related variables in the last two years are considerably more variable than other categories such as global macroeconomic indexes and international exchange rates. Bitcoin prices and volatilities have nearly doubled over the past two years. In addition, Blockchain data exhibit a significant increase in trading volume and size per a block and a huge reduction in miner's profit and the hash rate.

On the other hand, there is little difference between the most recent two years and the overall range in the volatility of the global exchange rate market as well as the growth of the global macroeconomic market economy over the past two years is much smaller than that of Bitcoin. These results provide empirical evidence for the fact that the recent volatility in Bitcoin prices stems mostly from the Blockchain information directly involved in supply and demand of Bitcoin and not from other macro-financial markets.

V. Experimental Results

A. Structure of the experiment

Most of the previous studies have focused on either modeling Bitcoin price without considering its relationship to Blockchain information or identifying only its "linear" relationship to macroeconomic factors. The present study attempts to overcome these limitations by employing a Bayesian NN model that can investigate nonlinear influences of each relevant feature of input variables, the Blockchain information, and macroeconomic factors, on Bitcoin price formation. To this end, we first train a Bayesian NN to model Bitcoin price formation using given above-mentioned relevant features of the process. We have evaluated Bayesian NN in terms of training and test errors by using the representative non-linear methodologies, SVR, and the linear regression model as the benchmark methods.

Next, we develop a prediction model of the near-future price of Bitcoin after modeling the entire process. We configure forecasting models by *the rollover framework*, which is generally applied to portfolio theory. Rollover strategy is known as rolling a position forward which is closing out an old position and establishing a new position in a contract of the portfolio with a long time to maturity. In our experiments, the trained machine is closing out an old information and acquiring new data according to the rollover framework over time. Figure 3 shows a schematic rollover strategy employed in our empirical studies. At the initial training step, the machine is learned with N_{train} training data, and the prediction performance is measured using N_{test} test data. Next, after $t' - t$ time from time t, the machine is trained using again the N_{train} data from time t' to update old learning data, and the performance of N_{test} test data is thereafter measured. The machine is trained through the entire range in this way, and the average performance of prediction errors measured several times is evaluated.

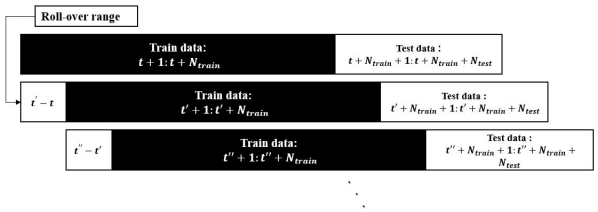


Fig. 3. the formation of the Blockchain

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779181, IEEE Access

TABLE II
SUMMARY STATISTICS OF THE DATA

Data category	Whole range		Recent 2 years		Data category	Whole range		Recent 2 years	
	mean	stdev.	mean	stdev.		mean	stdev.	mean	stdev.
Bitcoin price (USD)	458.32	606.2	901.96	804.0	**log of Bitcoin price (USD)**	5.04	1.92	6.52	0.71
volatility	10.75	25.06	21.83	38.88	**Trading volatility (BTC)**	$6.66*10^4$	$5.82*10^4$	$7.15*10^4$	$5.21*10^4$
Trading volatility (USD)	$3.36*10^7$	$6.59*10^7$	$6.96*10^7$	$9.77*10^7$	**Average block size**	$3.94*10^5$	$3.21*10^5$	$7.84*10^5$	$1.65*10^5$
Transactions per block	751.81	625.03	1507.61	389.58	**Median confirmation time**	9.15	3.59	10.21	3.44
Hash rate	$8.14*10^6$	$1.41*10^6$	$2.18*10^6$	$1.68*10^6$	**Difficulty**	$1.08*10^{11}$	$1.86*10^{11}$	$2.9*10^{11}$	$2.21*10^{11}$
Miners revenue (%)	2.7	2.17	1.04	0.42	**Miners revenue (USD)**	$1.36*10^6$	$1.38*10^6$	$2.16*10^6$	$1.57*10^6$
Confirmed transac. per day	$1.14*10^5$	$9.29*10^4$	$2.26*10^5$	$5.83*10^4$	**S&P 500**	1851.29	346.26	2169.8	166.84
Eurostoxx	2977.97	413.73	3208.97	235.1	**Dow Jones 30**	$1.64*10^4$	$2.59*10^3$	$1.87*10^4$	$1.71*10^3$
Nasdaq	4279.47	1029.08	5289.29	543.53	**Crudeoil**	73.53	25.21	45.23	5.98
SSE	2706.43	633.03	3140.39	223.83	**Gold**	1356.01	201.03	1218.72	78.67
VIX	16.02	5.09	15.12	4.65	**Nikkei225**	$1.50*10^4$	$3.87*10^3$	$1.82*10^4$	$1.46*10^3$
FTSE100	6444.86	549.54	6704.23	531.92	**USD/CNY**	6.37	0.25	6.65	0.2
USD/GBP	0.67	0.06	0.74	0.06	**USD/JPY**	102.36	14.67	112.4	6.44
USD/EUR	0.82	0.08	0.91	0.03	**USD/CHF**	0.95	0.04	0.99	0.02

Learning the machine through the rollover framework aims to validate the method of forecasting the next order of N_{test} test data from N_{train} training data. Given that the model employs time series in batch format, it is is faster and easier to learn than other sequential neural networks models, LSTM or RNN, and can reflect the flow of information that changes with time. The rollover framework can be used to implement semi-online prediction models to incorporate new information or shocks with short learning time.

B. Linear regression analysis

We first construct a linear model for analysis of Bitcoin price and address several critical issues in assumptions of the linear regression model. A basic assumption required for linear regression is *the model assumption* that linear relationships exist between response variables and independent variables [33]. Table III shows (linear) correlations between explanatory variables and response variables. Each column represents linear correlation coefficients of regressors for each response variable and the value in parentheses represents the results of t-test for the null hypothesis that there is no linear relationship between the two variables. We denote the null hypothesis-rejecting variables as bold, based on a p-value of 0.05, and presented a t-value because the p-value was as small as zero. We exclude the return as response variable because almost all values of correlation coefficients of each explanatory variable are not exceptionally significant for the return value of Bitcoin.

Next, we discuss the multicollinearity problem, which is often encountered in linear regression analysis. Several statistical problems are caused from the multicollinearity which is the situation that some regressors have a linear relationship with other regressors. It can cause undesirable regression analysis: very high R^2 for some coefficients that are not statistically significant and their t-statistics sensitive to data variation [33]. One of the prescriptions for dealing with multicollinearity is to do a linear regression except for variables with large VIF values, which is a sort of measure of the linear relationship between variables [33]. To remove redundant variables for preventing the collinearity problems, we eliminate several explanatory variables with large VIF values. Table IV shows VIF values of each explanatory variable. In this study, we have determined that the set of variables excluding linear relationships is suitable for linear regression analysis to avoid multicollinearity problem. We select 16 suitable discriminators after eliminating variables with large VIFs and perform linear regression analysis on Bitcoin log prices and log volatilities with these 16 discriminators. Removed variables include the following: transactions per a block, difficulty of the hash function, Nikkei225 index, S&P 500 index, Eurostoxx index, DOW30

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779181, IEEE Access

TABLE III
CORRELATION COEFFICIENTS AND (T-VALUES) BETWEEN THE RESPONSE AND INDEPENDENT VARIABLES.

Data category	return	price	log(price)	log(vol.)	Data category	return	price	log(price)	log(vol.)
Trading vol. (BTC)	**0.064 (2.987)**	**0.071 (3.315)**	**0.123 (5.772)**	**0.245 (11.769)**	**Trading vol. (USD)**	0.016 (0.745)	**0.777 (57.485)**	**0.474 (25.071)**	**0.683 (43.549)**
Avg. block size	0.001 (0.047)	**0.663 (41.246)**	**0.744 (51.857)**	**0.404 (20.569)**	**Trans. per block**	0.011 (0.512)	**0.647 (39.518)**	**0.715 (47.63)**	**0.39 (19.725)**
Median conf. time	0.04 (1.864)	**0.26 (12.54)**	0.018 (0.838)	**0.163 (7.694)**	**Hash rate**	0.025 (1.165)	**0.9 (96.16)**	**0.577 (32.902)**	**0.583 (33.419)**
Difficulty	0.024 (1.118)	**0.906 (99.686)**	**0.58 (33.159)**	**0.588 (33.856)**	**Miners revenue (%)**	-0.034 (-1.584)	**-0.34 (-16.838)**	**-0.51 (-27.613)**	**-0.24 (-11.514)**
Miners revenue (USD)	-0.015 (-0.699)	**0.92 (109.326)**	**0.76 (54.46)**	**0.625 (37.288)**	**Confirmed trans. per day**	0.008 (0.373)	**0.66 (40.915)**	**0.731 (49.891)**	**0.402 (20.447)**
S&P 500	-0.006 (-0.279)	**0.691 (44.52)**	**0.928 (116)**	**0.415 (21.243)**	**Eurostoxx**	-0.008 (-0.373)	**0.537 (29.647)**	**0.838 (71.523)**	**0.339 (16.782)**
Dow Jones 30	0.002 (0.093)	**0.746 (52.171)**	**0.916 (106.338)**	**0.454 (23.731)**	**Nasdaq**	-0.007 (-0.326)	**0.722 (48.599)**	**0.896 (93.973)**	**0.442 (22.948)**
Crudeoil	0.015 (0.699)	**-0.401 (-20.386)**	**-0.545 (-30.273)**	**-0.264 (-12.747)**	**SSE**	-0.018 (-0.838)	**0.27 (13.06)**	**0.408 (20.813)**	**0.184 (8.718)**
VIX	-0.051 (-2.378)	**-0.384 (-19.369)**	**-0.544 (-30.194)**	**-0.215 (-10.253)**	**Nikkei225**	-0.011 (-0.512)	**0.553 (30.911)**	**0.884 (88.067)**	**0.346 (17.175)**
FTSE100	0.016 (0.745)	**0.67 (42.033)**	**0.843 (72.987)**	**0.396 (20.085)**	**USD/CNY**	0.013 (0.605)	**0.572 (32.477)**	**0.355 (17.685)**	**0.331 (16.336)**
USD/GBP	0.019 (0.885)	**0.584 (33.506)**	**0.477 (25.276)**	**0.339 (16.782)**	**USD/JPY**	-0.018 (-0.838)	**0.38 (19.133)**	**0.819 (66.475)**	**0.244 (11.718)**
USD/EUR	-0.002 (-0.093)	**0.344 (17.062)**	**0.496 (26.603)**	**0.208 (9.904)**	**USD/CHF**	0.008 (0.373)	**0.266 (12.851)**	**0.341 (16.894)**	**0.164 (7.743)**
Gold	0.019 (0.885)	**-0.396 (-20.085)**	**-0.858 (-77.795)**	**-0.241 (-11.565)**					

TABLE IV
VIF VALUES OF EACH EXPLANATORY VARIABLE FOR DETECTING THE COLLINEARITY PROBLE

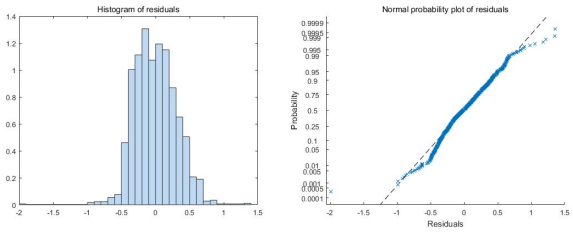
Data category	VIF	Data category	VIF	Data category	VIF	Data category	VIF
Trading vol. (BTC)	1.5688	**Trading vol. (USD)**	3.45327	**Avg. block size**	33.2689	**Trans. per block**	36.7642
Median conf. time	2.1306	**Hash rate**	122.3453	**Difficulty**	150.3203	**Miners revenue (%)**	2.4462
Miners revenue (USD)	8.2981	**Confirmed trans. per day**	48.1753	**S&P 500**	730.6197	**Eurostoxx**	41.9197
Dow Jones 30	402.9169	**Nasdaq**	304.5080	**Crudeoil**	22.8668	**SSE**	10.1965
Gold	21.4123	**VIX**	4.5702	**Nikkei225**	128.2556	**FTSE100**	51.7874
USD/CNY	20.3706	**USD/GBP**	45.355	**USD/JPY**	58.1390	**USD/EUR**	43.6925
USD/CHF	7.7059						

index, NASDAQ, and exchange rates of EUR and GBP. From these 16 regressors, we construct two linear models, one for the log price and one for the volatility of Bitcoin process. We then evaluate assumption fitness, say the *residual assumption* that residual terms are independently and identically distributed.

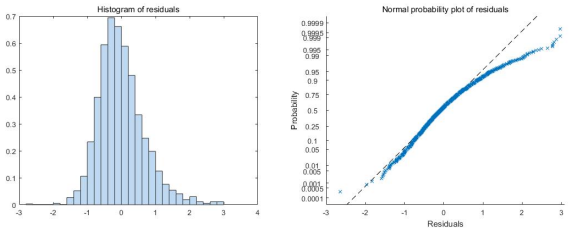
Finally, we generate histograms residuals of each model to verify the residual assumption by confirming it follows a normal distribution.

Figure 4 (a) & (b) show that the Bitcoin log price satisfies *the residual assumption* for linear regression: the histogram is bell-typed and symmetric and the QQ-plot shows a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779181, IEEE Access



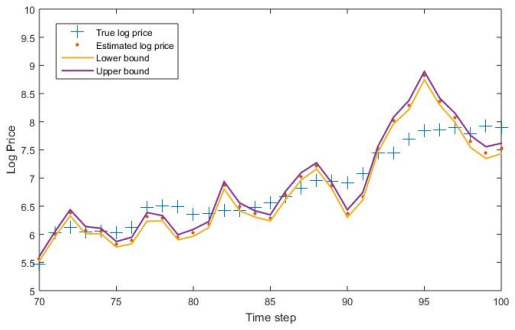
(a) Histogram for log-price (b) QQ plot for log-price



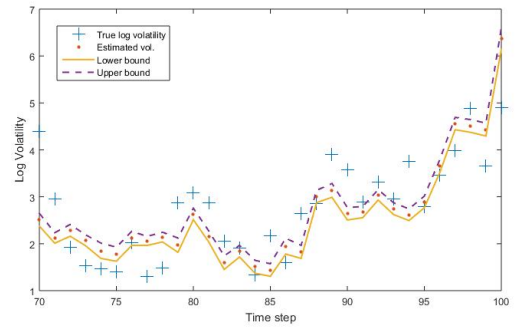
(c) Histogram for log-volatility (d) QQ plot for log-volatility

Fig. 4. Residual evaluations for (a) Histogram, (b) Normal probability (QQ) plot of the Bitcoin log price, and (c) Histogram, (d) Normal probability (QQ) plot of the Bitcoin log volatility

similar pattern with the normal distribution. By contrast, Figure 4 (c) & (d) show that residuals of the linear model for log volatility of Bitcoin do not follow a normal distribution with a positive-skewed histogram. Time series of log volatility of Bitcoin is therefore unsuitable for linear analysis except for the log price of Bitcoin due to the violation of each assumption.



(a) Bitcoin log-price



(b) Bitcoin log-volatility

Fig. 5. Prediction results of (a) the Bitcoin log price and (b) the Bitcoin log volatility

Each linear model trained from a random 85% of whole data are disparate from true log prices or log volatilities. Figure 5 demonstrates that predicted log prices (volatilities) and a confidence interval of most recent 30 test data, implying the unsuitability of the linear model in predicting the time series of Bitcoin price. Figure 5 shows that most true values are out of the confidence interval of the linear model. This means that the learned linear model does not make an adequate prediction of the output value albeit in predicting trends in little.

C. Results of Bitcoin price formation

We next perform time series analysis of Bitcoin prices using a BNN model and compare with the benchmark models, which are the linear regression and the SVR model. A total of 25 explanatory variables belonging to three categories are employed as inputs for BNN learning. We also address another input set that comprises 16 input variables by eliminating several unimportant variables as mentioned in the previous subsection. We consider two response variables, log price of Bitcoin and volatility of Bitcoin price, because extremely high volatility is an important feature of Bitcoin. In general, volatility is a significant variable assessed equally to the value of an option in economic analysis. We use log-scaled values of both output response variables to account for the large difference between Bitcoin value in the early period and its most recent value.

We train the BNN model through 10-fold cross-validation. To mitigate the effect of how to divide the data, we repeated hold-out validation steps where $\frac{9}{10}N$ training data and $\frac{1}{10}N$ test data, given the total number of the data is N. Where performances of each trained model are measured by root mean square error (RMSE) and mean absolute percentage error (MAPE). Definitions of each evaluation criteria are as followings:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{8}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{9}$$

where N is the number of samples, y_i is the i-th true objective value, and $\hat{y}_i$ is the i-th estimated value.

TABLE V
TRAINING ERROR FOR THE BITCOIN PRICE FORMATION

Response var.		Log price		Log volatility	
Num. of Input var.		26	16	25	16
Linear Regression	RMSE	-	0.0913	-	0.4595
	MAPE	-	0.0681	-	0.5905
Bayesian NN	RMSE	0.0031	0.0047	0.1612	0.1717
	MAPE	0.0119	0.0148	0.3314	0.3512
Support vec. Regression	RMSE	0.1453	0.1434	0.3810	0.3939
	MAPE	0.0325	0.0322	0.5411	0.6293

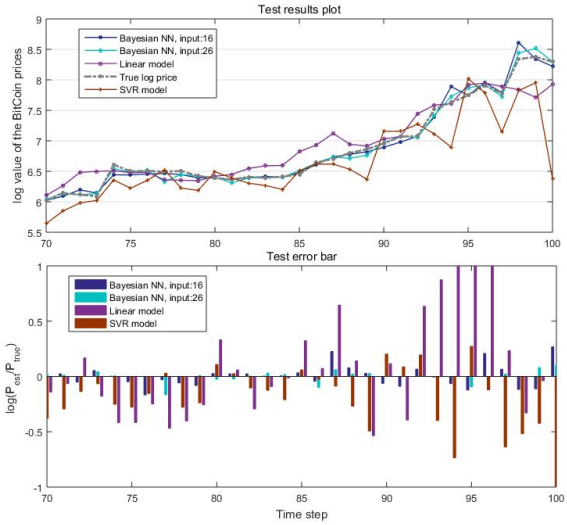
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779181, IEEE Access

TABLE VI
TEST ERROR FOR THE BITCOIN PRICE FORMATION

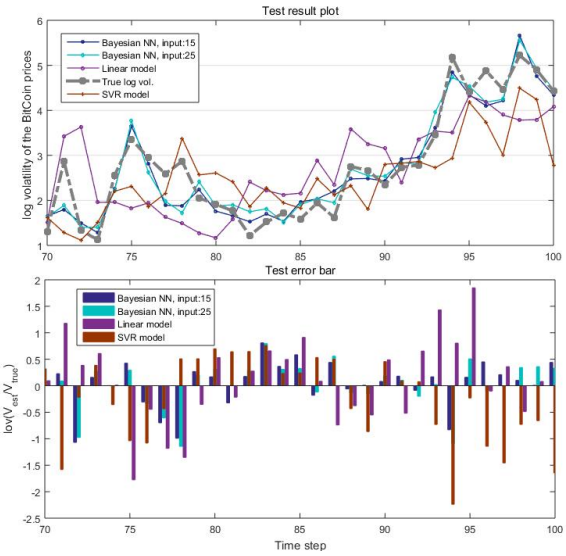
Response var.		Log price		Log volatility	
Num. of Input var.		26	16	25	16
Linear Regression	RMSE	-	0.0935	-	0.4823
	MAPE	-	0.0712	-	0.6263
Bayesian NN	RMSE	0.0039	0.0069	0.2546	0.2325
	MAPE	0.0138	0.0180	0.5090	0.5222
Support vec. Regression	RMSE	0.3201	0.2742	0.5487	0.5297
	MAPE	0.0428	0.0404	0.7232	0.8629

Table V and VI summarize results of training errors and test errors, respectively. We observe that BNN models outperform other models in terms of RMSE and MAPE for predicting the log price of Bitcoin. Log price of Bitcoin is learned exceptionally by the BNN model with training and test error of around 1% MAPE. In the case of log volatility, the prediction error of log volatility in the test phase is slightly larger than that in the training phase. BNN model is more reliable for describing the process of log volatility than other benchmark models. After eliminating redundant variables from linear correlation analysis, the error value is relatively small when all 26 input variables are considered instead of the abridged 16 input variables. This condition implies that removed variables may explain nonlinear relationships to adequately account for response variables. SVR model shows poor performances in both training and test phase. From this results, we can confirm that Bayesian NN is better suited for the Bitcoin time series analysis than SVR albeit in they are included the same nonparametric model.

Figure 6 shows the values of estimated response variables for the recent 30 test input data according to time indexes. We observe that the recent volatile tendency is well expressed in terms of explanatory input variables. The case of log price presents a tendency for underestimation when price rises and overestimation when the price falls. In the case of the log price, we can see that all models predict the actual tendency of the price to some extent. On the other hands, in terms of error size, it is confirmed that other models are larger than that of Bayesian neural networks. There is no tendency of over- or under-estimate in all models. Bayesian neural networks tended to predict consistent trends regardless of the number of inputs. In the case of volatility, the Bayesian NN model predicts better the direction of volatility than other benchmark models, and neither of the four models tends to over or under-estimate.



(a) Bitcoin log-price



(b) Bitcoin log volatility

Fig. 6. Test result plot of (a) the Bitcoin log price and (b) the Bitcoin log volatility

D. Prediction results under the rollover

Finally, we provide prediction results of the trained BNN under the rollover framework. Rollover framework physically excludes old preceding data to reflect that the previous information shrinks as training is repeated. The construction method of the model in this subsection is fundamentally different from that of the previous subsection. In the previous subsection, we have extracted part of the entire data for training purpose, assuming that we have all data for the entire time range. Although the method in the previous section is adequate to assess how well the model has learned for the whole data, it is not appropriate to predict future outcome from the historical data.

We train the machine using data obtained 200 days before the present day and predict the current day's price from the trained machine under the rollover framework. Given that future data are not considered in the training phase, we can infer that prediction performance may be inferior to that of the previous subsection. Table VII presents prediction error for Bitcoin price under the rollover framework. We note that overall performance is slightly poor compared

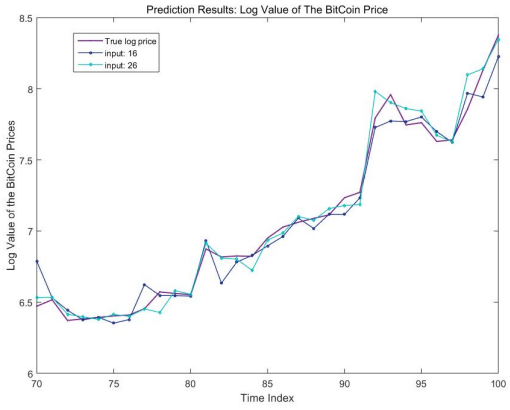
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779181, IEEE Access

with the model construction in the previous subsection. Nevertheless, prediction result for the log price of Bitcoin still maintains low error rates. By contrast, prediction errors are almost doubled for log volatility outputs. Figure 7 shows

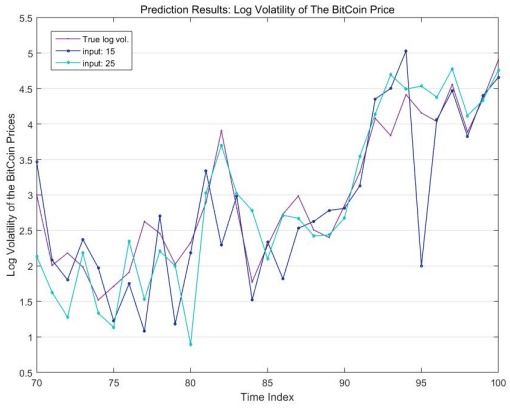
TABLE VII
PREDICTION ERROR FOR THE BITCOIN PRICE UNDER THE ROLLOVER FRAMEWORK

Response var.	Num. of input var.	RMSE	MAPE
log price	26	0.0256	0.0198
	16	0.0244	0.0200
log volatility	25	0.5750	0.8992
	15	0.5114	0.6302

plots of prediction results for the log price and log volatility of Bitcoin. We show that log price is relatively well explained based on the employed input variables and during sudden fluctuations. In the case of log volatility, the discrepancy between true volatility and predicted volatility is relatively large, but directionality is well approximated. In summary, the learned BNN models can effectively describe the recent highly volatile Bitcoin price process and the price in the entire range.



(a) Log value of the Bitcoin price



(b) Log volatility of the Bitcoin price

Fig. 7. Prediction results of (a) the log value of the Bitcoin price and (b) the log volatility of the Bitcoin price.

VI. CONCLUSION

Bitcoin is a successful cryptocurrency, and it has been extensively studied in fields of economics and computer science. In this study, we analyze the time series of Bitcoin price with a BNN using Blockchain information in addition to macroeconomic variables and address the recent highly volatile Bitcoin prices.

Given the data of the entire time range, experimental results show that the BNN model learned with the selected features effectively describes processes of Bitcoin log price and log volatility. Adoption of rollover framework experimentally demonstrates the predictive performance of BNN is better than other benchmark methods on log price and volatility processes of Bitcoin.

Through the empirical analysis, we have confirmed that the BNN model describes the fluctuation of Bitcoin up to August 2017, which is relatively recent. Unlike other benchmark models that fail directional prediction, the BNN model succeeded in relatively accurate direction prediction. From these experimental results, the BNN model is expected to have similar performance in more recent data. As the variation of Bitcoin process gets attention, it is expected that the expansion and application of the BNN model would be effective for the analysis and prediction of the Bitcoin process.

Investigating nonlinear relationships between input functions based on network analysis can explain analysis of Bitcoin price time series. Variability of Bitcoin must be modeled and predicted more appropriately. This goal can be achieved by adopting other extended machine learning methods or considering new input capabilities related to the variability of Bitcoin. Such study will contribute to rich Bitcoin time series analysis in addition to existing Bitcoin studies.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2016R1A2B3014030).

REFERENCES

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.
[2] A. H. Dyhrberg, "Bitcoin, gold and the dollar–a garch volatility analysis," *Finance Research Letters*, vol. 16, pp. 85–92, 2016.
[3] P. Katsiampa, "Volatility estimation for bitcoin: A comparison of garch models," *Economics Letters*, vol. 158, pp. 3–6, 2017.
[4] A. F. Bariviera, M. J. Basgall, W. Hasperué, and M. Naiouf, "Some stylized facts of the bitcoin market," *Physica A: Statistical Mechanics and its Applications*, vol. 484, pp. 82–90, 2017.
[5] J. Chu, S. Nadarajah, and S. Chan, "Statistical analysis of the exchange rate of bitcoin," *PloS one*, vol. 10, no. 7, p. e0133678, 2015.
[6] A. Urquhart, "The inefficiency of bitcoin," *Economics Letters*, vol. 148, pp. 80–82, 2016.
[7] S. Nadarajah and J. Chu, "On the inefficiency of bitcoin," *Economics Letters*, vol. 150, pp. 6–9, 2017.
[8] A. H. Dyhrberg, "Hedging capabilities of bitcoin. is it the virtual gold?" *Finance Research Letters*, vol. 16, pp. 139–144, 2016.
[9] E. Bouri, P. Molnár, G. Azzi, D. Roubaud, and L. I. Hagfors, "On the hedge and safe haven properties of bitcoin: Is it really more than a diversifier?" *Finance Research Letters*, vol. 20, pp. 192–198, 2017.
[10] E.-T. Cheah and J. Fry, "Speculative bubbles in bitcoin markets? an empirical investigation into the fundamental value of bitcoin," *Economics Letters*, vol. 130, pp. 32–36, 2015.

[11] L. Kristoufek, "Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era," *Scientific reports*, vol. 3, p. 3415, 2013.

[12] ——, "What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis," *PloS one*, vol. 10, no. 4, p. e0123923, 2015.

[13] P. Ciaian, M. Rajcaniova, and d. Kancs, "The economics of bitcoin price formation," *Applied Economics*, vol. 48, no. 19, pp. 1799–1815, 2016.

[14] S. McNally, "Predicting the price of bitcoin using machine learning," Ph.D. dissertation, Dublin, National College of Ireland, 2016.

[15] I. Madan, S. Saluja, and A. Zhao, "Automated bitcoin trading via machine learning algorithms," 2015.

[16] R. Gençay and M. Qi, "Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging," *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 726–734, 2001.

[17] R. J. Barro, "Money and the price level under the gold standard," *The Economic Journal*, vol. 89, no. 353, pp. 13–33, 1979.

[18] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder, *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton University Press, 2016.

[19] J. Pati, B. Kumar, D. Manjhi, and K. Shukla, "A comparison among arima, bp-nn and moga-nn forsoftware clone evolution prediction," *IEEE Access*, 2017.

[20] A. R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel, and A. S. Malik, "Machine learning framework for the detection of mental stress at multiple levels," *IEEE Access*, vol. 5, pp. 13 545–13 556, 2017.

[21] Y. Zhou, J. Yu, and X. Wang, "Time series prediction methods for depth-averaged current velocities of underwater gliders," *IEEE Access*, vol. 5, pp. 5773–5784, 2017.

[22] J. Yao and C. L. Tan, "Time dependent directional profit model for financial time series forecasting," in *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, vol. 5. IEEE, 2000, pp. 291–296.

[23] P. Wang, "Pricing currency options with support vector regression and stochastic volatility model with jumps," *Expert Systems with Applications*, vol. 38, no. 1, pp. 1–7, 2011.

[24] A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, "Support vector regression with chaos-based firefly algorithm for stock market price forecasting," *Applied soft computing*, vol. 13, no. 2, pp. 947–958, 2013.

[25] F. Kaytez, M. C. Taplamacioglu, E. Cam, and F. Hardalac, "Forecasting electricity consumption: a comparison of regression analysis, neural networks and least squares support vector machines," *International Journal of Electrical Power & Energy Systems*, vol. 67, pp. 431–438, 2015.

[26] T. Xiong, Y. Bao, and Z. Hu, "Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting," *Knowledge-Based Systems*, vol. 55, pp. 87–100, 2014.

[27] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[28] M. Minsky and S. Papert, "Perceptrons." 1969.

[29] B. Efron and R. Tibshirani, "Improvements on cross-validation: the 632+ bootstrap method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 548–560, 1997.

[30] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American statistical association*, vol. 78, no. 382, pp. 316–331, 1983.

[31] P. Jonathan, W. J. Krzanowski, and W. McCarthy, "On the use of cross-validation to assess performance in multivariate prediction," *Statistics and Computing*, vol. 10, no. 3, pp. 209–229, 2000.

[32] D. van Wijk, "What can be expected from the bitcoin," *Erasmus Universiteit Rotterdam*, 2013.

[33] D. N. Gujarati and D. C. Porter, "Essentials of econometrics," 1999.

Huisu Jang received the B.Sc., and the M.S. degree in industrial engineering from the Seoul National University (SNU), South Korea, in 2013, and 2015, where she is currently pursuing the ph.D. degree with the department of industrial engineering. Her research interests include statistical machine learning, time series analysis, and computational finance.

Jaewook Lee is a professor and chair in the Department of Industrial Engineering at Seoul National University, Seoul, Korea. He received the B.S. degree in mathematics from Seoul National University, and the Ph.D. degree in applied mathematics from Cornell University in 1993 and 1999, respectively. His research interests include machine learning, neural networks, global optimization, and their applications to data mining and financial engineering.