KIPS Trans. Softw. and Data Eng.
Vol.11, No.1 pp.19~28
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.1.19

Understanding the Association Between Cryptocurrency Price Predictive Performance and Input Features

Jaehyun Park† · Yeong-Seok Seo††

ABSTRACT

Recently, cryptocurrency has attracted much attention, and price prediction studies of cryptocurrency have been actively conducted. Especially, efforts to improve the prediction performance by applying the deep learning model are continuing. LSTM (Long Short-Term Memory) model, which shows high performance in time series data among deep learning models, is applied in various views. However, it shows low performance in cryptocurrency price data with high volatility. Although, to solve this problem, new input features were found and study was conducted using them, there is a lack of study on input features that drop predictive performance. Thus, in this paper, we collect the recent trends of six cryptocurrencies including Bitcoin and Ethereum and analyze effects of input features on the cryptocurrency price predictive performance through statistics and deep learning. The results of the experiment showed that cryptocurrency price predictive performance the best when open price, high price, low price, volume and price were combined except for rate of closing price fluctuation.

Keywords : LSTM, Deep Learning, Input Feature, Cryptocurrency, Price Prediction, Data Analysis

암호화폐 종가 예측 성능과 입력 변수 간의 연관성 분석

박 재 현† · 서 영 석††

요 약

최근 암호화폐가 많은 주목을 받음에 따라 암호화폐의 종가 예측 연구들이 활발히 진행되고 있다. 특히 딥 러닝 모델을 적용시켜 예측 성능을 높이려는 연구들이 지속되고 있다. 딥 러닝 모델 중 시계열 데이터에서 높은 예측 성능을 보이는 LSTM (Long Short-Term Memory) 모델이 다각도로 응용되고 있으나 변동성이 큰 암호화폐 종가 데이터에서는 낮은 예측 성능을 보인다. 이를 해결하기 위해 새로운 입력 변수를 찾아내고, 이를 사용하는 종가 예측 연구가 수행되고 있다. 그러나 딥 러닝 기반의 암호화폐 종가 예측에 사용되는 데이터들의 각 입력 변수들이 예측 성능에 미치는 영향력이나 학습에 효율적인 입력 변수들의 조합에 관한 연구 사례가 부족한 실정이다. 따라서 본 논문에서는 Bitcoin과 Ethereum을 포함한 6가지 암호화폐의 최근 동향 자료를 수집하였고, 통계와 딥 러닝을 통해 입력 변수들이 암호화폐 종가 예측에 미치는 영향력을 분석한다. 실험 결과 모든 암호화폐의 종가 예측 성능 평가에서 종가 변동률을 제외한 개장가, 고가, 저가, 거래량, 종가를 조합했을 때 가장 우수한 성능을 보였다.

키워드 : LSTM, Deep Learning, 입력 변수, 암호화폐, 가격 예측, 데이터 분석

1. 서 론

Satoshi Nakamoto가 2008년 Bitcoin을 처음 소개한 이래로 수많은 암호화폐가 등장했으며, 현재 관련 연구도 활발히 진행되고 있다. 암호화폐 종가 예측 연구에서는 딥 러닝을 응용해 다양한 학습 데이터를 다양한 관점에서 분석하거나, RNN (Recurrent Neural Network) 모델의 느린 학습 속도와 “vanishing gradient” 문제를 해결한 시계열 데이터에서 우수한 예측 성능을 보이는 LSTM (Long Short-Term Memory) 모델을 사용해 암호화폐 종가를 예측하는 방법이 최근까지 대표적인 연구 사례이다[1,2].

그러나 딥 러닝을 통한 암호화폐 종가 예측은 급격한 종가 변동에 대한 패턴이 존재할 경우 낮은 예측 성능을 보인다[3]. 이러한 문제를 해결하기 위해 종가와 관련된 입력 변수뿐만 아니라 새로운 입력 변수를 찾아내고, 이를 사용하는 종가 예측 연구가 수행되었다[1]. 그러나 암호화폐 종가 예측에

※ 이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2020R1I1A3073313).
※ 이 논문은 2021년 한국정보처리학회 춘계학술발표대회의 우수논문으로 “LSTM 모델의 하이퍼 파라미터가 암호화폐 가격 예측에 미치는 영향 분석”의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 영남대학교 컴퓨터공학과 석사과정
†† 종신회원 : 영남대학교 컴퓨터공학과 교수
Manuscript Received : June 25, 2021
First Revision : August 6, 2021
Accepted : August 11, 2021
* Corresponding Author : Yeong-Seok Seo(ysseo@yu.ac.kr)

효율적인 입력 변수 조합이나 각 입력 변수들이 미치는 긍정적 및 부정적 영향에 관한 연구는 부족하다.

따라서 본 논문에서는 Binance, Bitcoin, Cardano, Chainlink, Ethereum, Litecoin과 같은 대표적인 암호화폐의 최근 추세 시계열 데이터를 활용하여 각 입력 변수와 이들의 조합이 암호화폐 종가 예측에 미치는 영향력을 분석한다. 이를 위해 우리는 먼저 개장가, 고가, 저가, 거래량, 종가 변동률, 종가 총 6개의 각 입력 변수와 암호화폐 종가 간의 상관관계를 통계적인 분석으로 검토하였다. 그다음 입력 변수들의 다양한 조합을 학습한 딥 러닝 기반의 암호화폐 예측모델을 평가하여 각 입력 변수들의 영향력과 효율적인 조합에 대해 분석하였다.

새로운 입력 변수를 찾아내어 예측 성능을 향상시키는 기존 연구들과 달리 본 논문에서는 실험에 사용된 입력 변수에 대해 예측 성능을 악화시키는 입력 변수를 찾아 이를 제거하며, 가장 우수한 예측 성능을 보이는 입력 변수 조합을 찾는 데 기여할 것으로 예상된다.

이하, 1장 서론에 이어 본 논문의 구성은 다음과 같다. 2장에서는 예측 성능을 높이기 위해 새로운 입력 변수를 사용한 관련 연구들을 기술한다. 3장에서는 실험에서 사용한 하이퍼파라미터와 본 논문에서 수행될 실험에 대한 세부 정보 및 평가 기준과 같은 실험 설계를 설명한다. 다음 4장에서는 제시된 평가 기준으로 얼마나 예측 성능이 차이 나는지 실험 분석을 기술한다. 마지막으로 5장에서는 결론 및 향후 연구를 기술한다.

2. 관련 연구

암호화폐 종가 예측 성능을 높이기 위해 새로운 입력 변수를 찾아내기 위한 연구는 2015년부터 현재까지 계속해서 연구되고 있다[1]. 본 장에서는 암호화폐 종가 예측을 위한 선행 연구들에 대해 소개한다. 암호화폐 종가 예측 연구는 크게 3가지 소셜미디어, 커뮤니티 또는 뉴스 등에서 발생되는 사용자의 의견이나 감성을 이용한 방법, 시계열로부터 주파수를 분해하여 특징을 추출하는 EMD (Empirical Mode Decomposition)를 이용하는 방법, 기존에 존재하는 다양한 보조 지표를 활용하는 방법으로 구분된다[4-13].

2.1 자연어 처리 기반 종가 예측 연구

J. Zhigang의 연구는 고가, 저가, 개장가, 종가 데이터에 감성 분석을 위해 긍정/부정에 대한 정보를 입력 변수로 추가했다. 특정 일의 특정 암호화폐에 대해 주식 커뮤니티 사이트에서 대중들의 긍정적인 리뷰가 많은지 부정적인 리뷰가 많은지를 판단한 뒤, 이 정보를 입력 변수로 사용했다. 기존 입력 변수에 감성 분석 데이터를 추가했을 때 더 높은 예측 성능을 보인다는 것을 증명했다[4].

H. Yamamoto의 연구는 SNS 중 하나인 트위터에서 대중들이 아닌 인플루언서들의 리뷰를 사용하였다. 이전 주식 커뮤니티 사이트에서 대중들의 리뷰를 사용한 연구와 마찬가지로 긍정/부정 중 어떤 종류의 리뷰가 많은지를 판단한 뒤, 이를 입력 변수로 사용했다. 이 또한 더 높은 예측 성능을 보인다는 것을 증명했다[5].

H. Maqsood의 연구는 감성 분석을 위한 긍정/부정에 대한 정보뿐만 아니라 대중들의 중립적인 리뷰까지 사용했다. 긍정/부정에 대한 정보만 입력 변수로 사용했을 때보다 긍정/부정/중립에 대한 정보를 입력 변수로 사용했을 때 더 높은 예측 성능을 보인다는 것을 증명했다[6].

Z. Hu의 연구는 종가, 거래량 데이터에 뉴스 헤드라인을 사용했다. 뉴스 헤드라인에 긍정적인 뉴스 헤드라인이 많은지 부정적인 뉴스 헤드라인이 많은지를 판단한 뒤, 이 정보를 입력 변수로 사용했다. 기존 입력 변수에 이 데이터를 추가했을 때 더 높은 예측 성능을 보인다는 것을 증명했다[7].

2.2 EMD 알고리즘 기반 종가 예측 연구

Y. Xuan의 연구는 고가, 저가, 개장가, 종가 데이터에 대해 EMD 알고리즘을 통해 변동성이 작은 데이터로 변환하는 전처리 과정을 추가했다. EMD 알고리즘은 신호 또는 그래프 데이터에서 고주파 성분에서 저주파 성분까지 각각 추출하는 역할을 한다. 이 주파수 성분들은 변동성이 작은 데이터이기 때문에 기존 입력 변수와 이 주파수 성분들을 같이 사용했을 때 더 높은 예측 성능을 보인다는 것을 증명했다[8].

R. Hadi의 연구는 CEEMD (Complementary Ensemble Empirical Mode Decomposition) 알고리즘을 사용했다. 이 알고리즘은 EMD 알고리즘의 'Mode Mixing'이라는 문제점이 개선된 형태이다[9]. EMD 알고리즘을 통해 나온 주파수 성분보다 CEEMD 알고리즘을 통해 나온 주파수 성분을 사용할 때 더 높은 예측 성능을 보인다는 것을 증명했다[10].

2.3 보조 지표 기반 종가 예측 연구

Y. Li의 연구는 고가, 저가, 개장가, 종가 데이터 대신 주식 그래프 중 하나인 캔들스틱 차트를 강화학습 모델에 사용했다. 강화학습 모델에서는 종가 관련 데이터 대신 캔들스틱 차트를 사용할 때 더 높은 예측 성능을 보인다는 것을 증명했다[11].

BS. Lin의 연구는 고가, 저가, 개장가, 종가 데이터에 RSI (Relative Strength Index)를 추가했다. RSI는 주식의 기술적 분석에 사용되는 종가 보조 지표이다. 이 지표를 함께 사용했을 때 더 높은 예측 성능을 보인다는 것을 증명했다[12].

P. Oncharoen의 연구는 종가 데이터에 Stochastic Oscillator를 추가했다. Stochastic Oscillator는 최근 N 일간의 고가와 저가 범위에서 현재 종가의 위치를 표시한 주식의 기술적 분석에 사용되는 보조 지표이다. 이를 함께 사용했을 때 더 높은 예측 성능을 보인다는 것을 증명했다[13].

Z. Li의 연구는 종가 데이터에 Technical Indicator를 추가했다. Technical Indicator는 종가 또는 거래량에 대한 금융 시장 방향을 예측하는 보조 지표이다. 이 지표를 함께 사용

했을 때 더 높은 예측 성능을 보인다는 것을 증명했다[14].

이처럼 종가 예측 성능을 향상시키는 입력 변수들에 관한 연구는 많지만, 종가 예측 성능을 하락시키는 입력 변수에 관한 연구는 부족하다. 따라서 본 논문에서는 이러한 입력 변수에 대해 초점을 맞추기로 했다.

3. 실험 설계

3.1 실험 데이터

실험 데이터로는 현재 가장 널리 알려진 Bitcoin과 Ethereum, 암호화폐 순위가 10위 이내 및 암호화폐가 본격적으로 떠오르기 시작한 2017년 하반기부터 데이터를 가지고 있으면서 보안성과 확장성에서 장점이 있는 Binance와 Cardano와 Chainlink 그리고 Litecoin까지 총 6가지의 암호화폐를 사용한다[15-20].

Bitcoin은 2008년 10월 Satoshi Nakamoto가 소개한 최초의 암호화폐이다. 중앙은행이 없어 개인들 간에 자유로운 거래가 가능하다. 또한, 거래 내역을 SHA-256 기반의 함수로 암호화를 한 뒤 블록체인 기술을 기반으로 분산하여 저장하기 때문에 해킹의 위험이 적다[15]. Ethereum은 2015년 7월 Vitalik Buterin이 소개한 암호화폐이다. Bitcoin과 달리, 거래 내용뿐만 아니라 계약서와 같은 추가 정보까지 기록할 수 있으며, IoT (Internet of Things)에 적용할 수 있는 특징 때문에 큰 주목을 받고 있다[16].

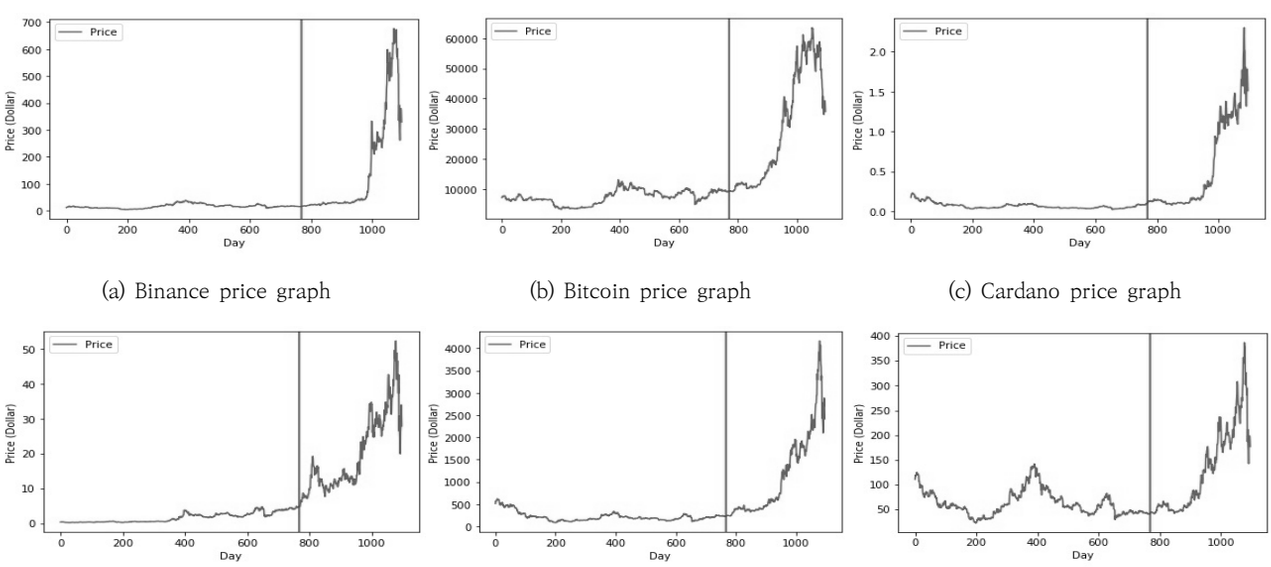
Binance는 2017년 7월 Changpeng Zhao가 소개한 암호화폐이다. Ethereum 기반의 표준 사양인 ERC-20을 사용하는 다른 암호화폐들과 달리 자체적인 사양을 사용한다는 특징이 있다[17]. Cardano는 2017년 9월 Charles Hoskinson이 소개한 암호화폐이다. 다른 암호화폐와 달리 확장성에서 용이하며, 연간 소모 전력 또한 Bitcoin의 0.01% 미만일 정도로 개선되었다[18].

Chainlink는 2017년 9월 Sergey Nazarov가 소개한 암호화폐이다. 블록체인 내부의 데이터와 블록체인 외부의 데이터를 연결하는 블록체인 미드웨어 역할을 한다는 특징이 있으며, 이러한 특징 때문에 NFT 분야에서 큰 주목을 받고 있다[19]. Litecoin은 2011년 10월 Charlie Lee가 소개한 암호화폐이다. 암호화폐에 대한 거래 내용을 모두 저장하는 기존 암호화폐와 달리 모든 거래 내용을 종합한 뒤 최종 결과만 업로드하는 라이트닝 네트워크 솔루션을 도입해 처리 속도 문제를 개선했다[20].

Fig. 1은 암호화폐별 종가 그래프이다. 각 그래프에 대해 가로축은 날짜, 세로축은 해당 날짜의 암호화폐 종가를 나타낸 것이다. 암호화폐가 본격적으로 주목을 받은 시기는 2017년 후반기이다. 이에 따라 기록된 데이터 개수가 부족하므로 현재 구할 수 있는 데이터 중 2018년 05월 28일부터 2021년 05월 28일까지 1,097일간의 데이터를 사용하며 데이터 개수 중 70%인 769개는 훈련 데이터, 나머지 30%인 328개는 검증 데이터로 사용한다. Fig. 1에 대해 빨간 수직선을 기준으로 좌측이 훈련 데이터, 우측이 검증 데이터이다.

3.2 입력 변수

본 연구에서는 보다 신뢰성 있는 연구 결과 도출을 위해 글로벌 암호화폐 거래소에서 거래 점수 상위 20개인 Coinbase Exchange, Huobi Global, Kraken, FTX, Bitfinex, KuCoin, Bithumb, Gate.io, Binance.US, Bitstamp, Gemini, bitFlyer,



(a) Binance price graph (b) Bitcoin price graph (c) Cardano price graph

(d) Chainlink price graph (e) Dogecoin price graph (f) Ethereum price graph

Fig. 1. Cryptocurrency Price Graph

Table 1. Value of LSTM Hyper-parameter

Input Feature	Explanation
Change %	The rate of increase and decrease on the day-to-day closing price
High	The price at the time when the price was the highest on the day
Low	The price at the time when the price was the lowest on the day
Open	The price at the time the day was first opened
Price	The price at the time of closing the day
Volume	The amount of transactions that occurred on the same day

Table 2. Value of LSTM Hyper-parameter

Hyper-parameter	Value
intput_dim	6
hidden_dim	30
num_layers	1
optimizer	Adam
loss function	MSE
learning rate	0.001
epoch	100

Coinone, Poloniex, Bitrex, Liquid, OKEx, FTX US, Korbit에서 제공하고 있는 모든 입력 변수들인 종가 변동률 (Change %), 고가 (High), 저가 (Low), 개장가 (Open), 종가 (Price), 거래량 (Volume) 총 6가지의 입력 변수를 확보하여 실험에 사용하였으며, 각 입력 변수에 대한 설명은 Table 1과 같다. 편의상 본 논문에서 종가 변동률은 'C', 고가는 'H', 저가는 'L', 개장가는 'O', 종가는 'P', 거래량은 'V'로 표기한다[21].

3.3 사용 모델 및 하이퍼 파라미터

실험에 사용한 딥 러닝 모델은 Tensorflow 2.0의 LSTM 모델을 사용했으며, 과거 일주일간의 데이터를 통해 현재의 종가를 예측한다. 또한, 하이퍼 파라미터는 기존 암호화폐 예측 연구와 동일하게 설정했으며 Table 2는 실험에 사용된 딥 러닝 모델에 대한 하이퍼 파라미터와 그 값이다[22].

고가, 저가, 개장가, 종가, 거래량, 종가 변동률 총 6가지를 입력 변수로 사용한다. 내부 node의 수는 30개이며, hidden layer 수는 1층으로 구성하였다. optimizer은 Adam optmimzer를, loss function은 MSE loss function을 사용했다. learning rate는 0.001이며 100번의 epoch만큼 학습한다.

3.4 평가 기준

1) 통계기반 분석

본 절에서는 입력 변수에 대한 관계성 검증을 위해 상관석과 신뢰성 검증을 위해 다중 공선성 분석 중 하나인 VIF (Variance Inflation Factor)을 설명한다.

Table 3. Value of Pearson Correlation Coefficient

Range	Explanation
-1.0 ~ -0.7	Strong negative linear relation
-0.7 ~ -0.3	Distinct negative linear relation
-0.3 ~ -0.1	Weak negative linear relation
-0.1 ~ +0.1	Negligible linear relationship
+0.1 ~ +0.3	Weak positive linear relation
+0.3 ~ +0.7	Distinct positive linear relation
+0.7 ~ +1.0	Strong positive linear relation

제시하는 평가 기준 각각은 단독적으로 사용되어 절대적인 기준이 될 수 없다. 따라서 2가지 기준을 모두 적용하여보다 더 융통성 있는 평가를 수행한다.

a) 상관분석

상관분석은 두 입력 변수 간에 어떤 선형적 또는 비선형적 관계가 있는지 분석하는 기법이다. 본 실험에서는 피어슨 상관계수를 사용하며, 두 변수에 대해 완전히 동일한 양의 선형관계를 가지면 1, 완전히 관계가 다르면 0, 완전히 동일한 음의 선형관계를 가지면 -1을 가진다. 값의 범위에 대한 설명은 Table 3과 같다[23].

b) 다중 공선성 분석

다중 공선성 분석은 각 입력 변수 간 상관관계를 분석하는 기법이다. 입력 변수 간 상관관계가 크면 다중 공선성이 존재함을 의미하며, 이로 인해 과대 적합을 발생시킬 수 있다. 본 실험에서는 VIF를 사용하며, VIF가 10 이상일 경우 다중 공선성이 있다고 판단한다[24].

2) 딥 러닝 기반 특징별 영향력 분석

본 절에서는 최적의 딥 러닝 예측 성능에 대한 검증을 위해 MSE (Mean Squared Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2를 이용한 딥 러닝 기반 암호화폐 종가 예측모델의 성능을 평가하기 위한 기준을 설명한다.

제시하는 평가 기준 각각은 단독적으로 사용되어 절대적인 기준이 될 수 없다. 따라서 4가지 기준을 모두 적용하여 보다 더 융통성 있는 평가를 수행한다.

a) MSE

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2 \quad (1)$$

MSE의 공식은 위의 식과 같다. 실제값 y_i에서 예측값 $\hat{y_i}$의 오차에 대해 제곱을 한 뒤, 그 합에 대한 평균값이다. 값이 작을수록 더 높은 예측 성능을 의미한다.

b) RMSE

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

RMSE의 공식은 위의 식과 같다. MSE 값에 제곱근을 한 값이다. MSE와 마찬가지로 값이 작을수록 더 높은 예측 성능을 의미한다.

c) MAE

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (3)$$

MAE의 공식은 위의 식과 같다. 실제값 y_i와 예측값 $\hat{y}_i$의 절대 오차에 대해 그 합에 대한 평균값이다. MSE, RMSE와 마찬가지로 값이 작을수록 더 높은 예측 성능을 의미한다.

d) R^2

R^2는 예측값이 실제값을 얼마나 나타내는지에 대한 척도이다. 실제값과 예측값의 차에 대한 평균을 사용하는 MSE, RMSE, MAE는 그 값이 실제값을 얼마나 반영하는지 알아보기에는 한계가 있으므로 R^2를 함께 사용한다. 값이 음수일 때는 그 값과 관계없이 예측값이 실제값을 전혀 반영하지 않는다는 것을 의미한다. 값이 양수일 때는 0%~100% 사이의 값을 가지며 그 값이 클수록 더 높은 예측 성능을 의미한다. 기존 연구에서 0.1%의 차이도 유의미하다고 증명했다[25]. 따라서 본 실험에서도 의미가 있다고 정의한다.

4. 실험 분석

4.1 통계기반 분석

본 절에서는 입력 변수들이 암호화폐의 종가에 미치는 영향력을 확인하기 위한 통계기반의 실험을 수행한다. 첫 번째 실험으로 입력 변수에 대해 피어슨 상관계수를 기반으로 한 상관분석을 수행하여 각 입력 변수 간의 관계성을 실험 및 검증한다. 두 번째 실험으로 종가 입력 변수와 나머지 입력 변수 간의 상관관계를 VIF를 기반으로 한 다중 공선성 분석 통해 실험 및 검증한다.

Fig. 2는 암호화폐별 입력 변수에 대한 상관계수를 통계분석 프로그래밍 언어 중 한 종류인 'R'을 통해 나타낸 결과이다. 각 표에 대해 여섯 번째 행은 종가에 대한 나머지 입력 변수 간의 피어슨 상관계수 값으로 좌측 1열을 기준으로 차례로 개장가, 고가, 저가, 거래량, 종가 변동률 간의 값이다. 모든 암호화폐에 대해 개장가, 고가, 저가는 종가에 대해 값의 편차가 크지 않고 거의 완벽한 선형적 관계를 보인다. 이로 인해 0.99에서 1.00 사이의 값을 가지고 있으며, 이를 통해 매우 강한 상관관계를 가지고 있음을 알 수 있다. 거래량과 종가 간의 상관계수는 각 암호화폐 별로 0.11에서 0.64 사이로 나타나며 종가형성에 전반적으로 적지 않은 영향력을 미침을 알 수 있다. 마지막으로 종가 변동률과 종가 간의 상관계수는 0.03에서 0.07 사이의 값을 가지고 있으며, 이를 통해 두 변수 간의 상관관계가 거의 없음을 알 수 있다.

Table 4는 종가 입력 변수에 대한 나머지 입력 변수 간의 VIF를 나타낸 것으로 10 이상이면 종가와 독립적이지 않음을 의미한다. 모든 암호화폐에 대해 개장가, 고가, 저가는 값이 10 이상으로 이 변수들은 상호 독립적이지 않다. 반면 모든 암호화폐에 대해 거래량과 종가 변동률은 값이 10 미만으로 이를 통해 입력 변수 간 다중 공선성이 존재하지 않고 상호 독립적인 변수임을 알 수 있다.

두 실험 결과를 종합해보았을 때 개장가, 고가, 저가는 종가에 대해 매우 강한 상관관계를 가지고 있으나 상호 간의 다중 공선성이 존재해 독립적인 변수가 아니다. 이처럼 독립적이지 않은 변수는 딥 러닝 기반 암호화폐 종가 예측모델의 학습에 있어 과대 적합을 발생시킬 가능성이 있다. 거래량의 경우 종가에 대해 무시할 수 없는 상관관계를 가지고 있어 예측 성능에 긍정적인 영향을 끼칠 수 있으며, 다중 공선성이 존재하지 않아 독립적인 변수임을 알 수 있다. 종가 변동률의 경우 종가에 대해 다중 공선성이 존재하지 않지만, 상관관계가 거의 존재하지 않아 종가 예측에 영향을 미치지 않고 오히려 부정적 영향을 끼칠 수 있다.

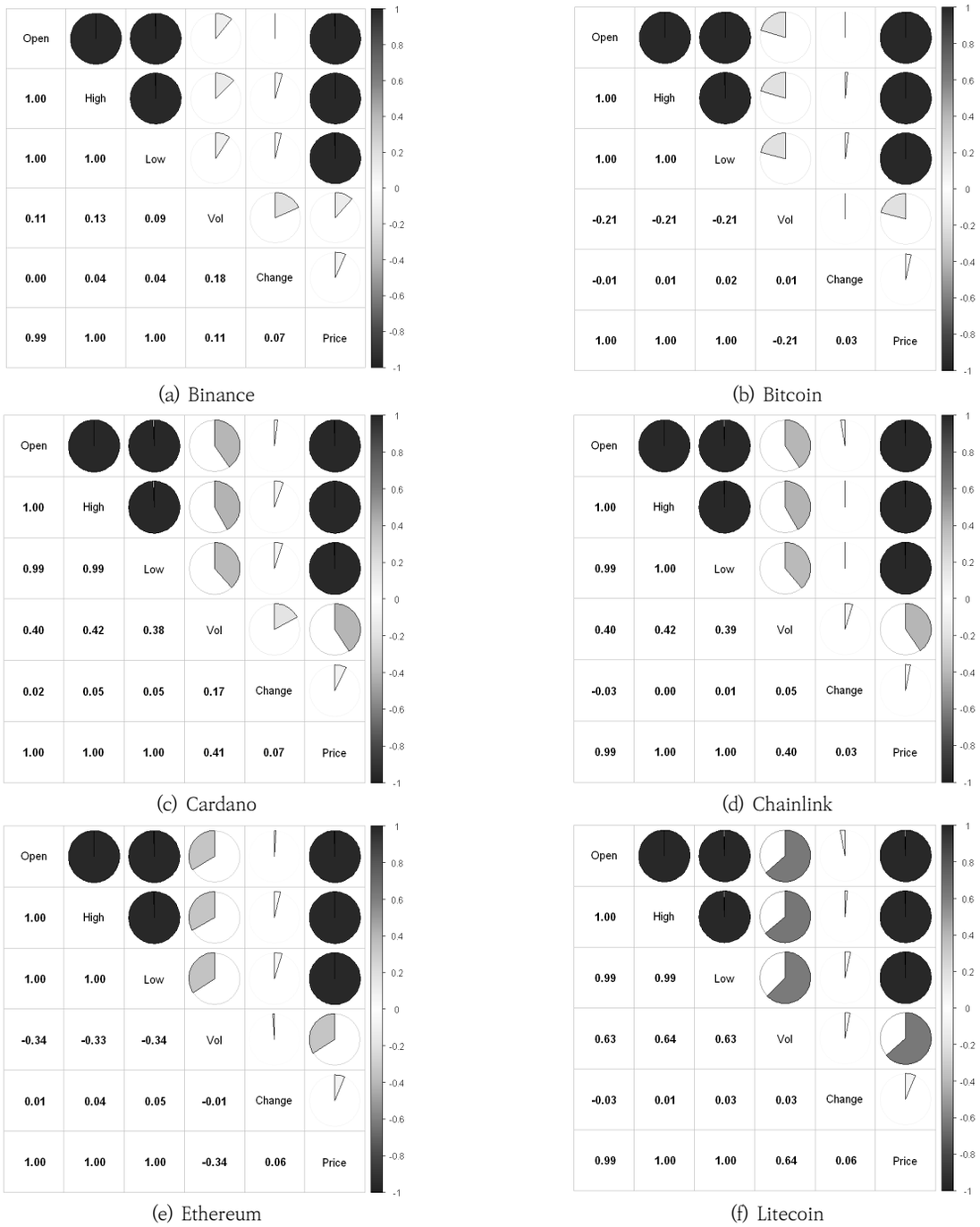
4.2 딥 러닝 기반 특징별 영향력 분석

본 절에서는 입력 변수들이 딥 러닝 기반 암호화폐 종가 예측에 미치는 영향력을 확인하기 위한 실험을 수행한다. 먼저 예측에 사용될 입력 변수의 개수를 1개에서 6개까지 늘려가면서 입력 변수를 조합한다. 각 조합에 대해 암호화폐별 가장 우수한 예측 성능을 보이는 조합을 딥 러닝 기반 암호화폐 종가 예측모델을 통해 검증한다. 또한, 통계기반 분석에서 부정적 영향을 끼칠 수 있을 것으로 예상되는 종가 변동률과 긍정적 영향을 끼칠 수 있을 것으로 예상되는 거래량과 과대적합을 발생시킬 가능성이 있는 개장가 그리고 고가, 저가가 실제 딥 러닝 기반 암호화폐 종가 예측에서도 동일한 영향을 끼치는지 실험 및 검증한다.

Table 5는 입력 변수의 개수에 따른 가장 우수한 딥 러닝 예측 성능을 보이는 입력 변수 조합, 입력 변수가 선택된 횟수, 그리고 평가 기준으로 사용된 R^2, MSE, RMSE, MAE를 나타낸 결과이다.

입력 변수의 개수가 1개인 경우 R^2를 기준으로 43.51에서 74.95 사이의 예측 성능을 보임을 알 수 있다. 단일 변수만 사용한 평가에서 고가 변수가 가장 채택빈도가 높았으며 종가 변동률의 경우 아예 선택되지 않았다. 그 외 종가와 저가 그리고 거래량이 1번씩 채택되었다.

입력 변수의 개수가 2개인 경우 R^2를 기준으로 74.09에서



(a) Binance (b) Bitcoin

(c) Cardano (d) Chainlink

(e) Ethereum (f) Litecoin

Fig. 2. Result of Correlation Analysis

Table 4. Result of VIF

		Input Feature				
		O	H	L	V	C
Cryptocurrency	Binance	480.7	373.1	175.9	1.2	1.7
	Bitcoin	1287.6	943.7	413.2	1.0	1.7
	Cardano	315.9	278.8	115.2	1.4	1.4
	Chainlink	449.5	459.9	131.0	1.4	1.4
	Ethereum	590.1	496.3	161.1	1.1	1.5
	Litecoin	494.8	370.2	104.1	1.7	2.2

86.63 사이의 예측 성능을 보이며, 모든 평가에서 각 암호화폐에 대해 입력 변수를 1개 사용한 경우보다 향상된 예측 성능을 보인다. 2개 변수 조합에서는 종가변수가 4개로 가장 많은 선택을 받았고 그 뒤를 이어 거래량이 3번 선택되었다. 반면 개장가는 단일 변수실험 때와는 달리 고가는 1번만 선택되었고 대신 개장가와 저가가 2번씩 선택되었다. 종가 변동률은 이번 실험에서도 선택받지 못했다.

입력 변수의 개수가 3개인 경우 R^2를 기준으로 84.24에서

Table 5. Comprehensive Analysis of Deep Learning Experiment at Each Cryptocurrency

Number of Input Features	Cryptocurrency	O	H	L	V	C	P	R^2	MSE	RMSE	MAE
One	Binance		o					73.47	4,540.47	67.38	60.52
	Bitcoin		o					61.35	52,459,035.47	7,242.86	6,337.47
	Cardano						o	68.44	0.14	0.37	0.26
	Chainlink				o			74.24	16.68	4.08	3.75
	Ethereum		o					43.51	160,414.67	400.52	324.68
	Litecoin			o				74.95	1,728.07	41.57	39.54
Sum		**0**	**3**	**1**	**1**	**0**	**1**				
Two	Binance	o					o	75.13	4,540.47	67.38	60.52
	Bitcoin			o			o	77.42	34,723,229.74	5,892.64	5,404.85
	Cardano				o		o	74.08	0.10	0.31	0.24
	Chainlink	o		o				86.57	9.29	3.05	2.61
	Ethereum		o		o			81.84	67,002.81	258.85	214.77
	Litecoin				o		o	86.63	1,042.26	32.28	26.86
Sum		**2**	**1**	**2**	**3**	**0**	**4**				
Three	Binance			o	o		o	84.21	3,349.40	57.87	50.52
	Bitcoin	o		o			o	86.07	26,567,365.16	5,154.35	4,641.14
	Cardano	o	o		o			86.21	0.04	0.20	0.16
	Chainlink	o		o			o	91.33	8.71	2.95	2.02
	Ethereum		o	o	o			87.45	56,683.99	238.08	190.51
	Litecoin	o			o		o	89.54	804.57	28.37	19.29
Sum		**4**	**2**	**4**	**4**	**0**	**4**				
Four	Binance	o	o	o			o	95.09	1,118.84	33.45	26.94
	Bitcoin		o	o	o		o	98.54	5,225,677.13	2,285.97	1,437.49
	Cardano	o	o		o		o	97.32	0.01	0.10	0.04
	Chainlink	o	o		o		o	93.89	7.86	2.80	1.72
	Ethereum	o	o		o		o	96.77	23,371.68	152.88	97.49
	Litecoin	o	o		o		o	97.89	169.26	13.01	7.08
Sum		**5**	**6**	**2**	**5**	**0**	**6**				
Five	Binance	o	o	o	o		o	98.45	516.06	22.72	10.59
	Bitcoin	o	o	o	o		o	99.09	2,944,935.92	1,716.08	1,129.76
	Cardano	o	o	o	o		o	98.46	0.01	0.09	0.04
	Chainlink	o	o	o	o		o	95.11	4.41	2.10	1.57
	Ethereum	o	o	o	o		o	98.66	12,032.20	109.69	62.78
	Litecoin	o	o	o	o		o	98.52	148.34	12.18	6.68
Sum		**6**	**6**	**6**	**6**	**0**	**6**				
Six	Binance	o	o	o	o	o	o	90.76	2,201.10	46.92	32.33
	Bitcoin	o	o	o	o	o	o	94.53	18,746,463.02	4,329.72	3,706.49
	Cardano	o	o	o	o	o	o	94.81	0.01	0.12	0.07
	Chainlink	o	o	o	o	o	o	93.47	8.62	2.94	1.77
	Ethereum	o	o	o	o	o	o	96.85	22,874.51	151.24	86.38
	Litecoin	o	o	o	o	o	o	97.15	210.85	14.52	8.02
Sum		**6**	**6**	**6**	**6**	**6**	**6**				

91.33 사이의 예측 성능을 보이며, 모든 평가 기준을 통해 각 암호화폐에 대해 입력 변수를 1개에서 2개 사용한 경우보다 조합별 실험에서 가장 높은 성능 향상 폭을 보인다. 3개 변수 조합에서는 고가와 종가 변동률을 제외한 모든 변수가 4번의 선택을 받았다. 고가는 2번의 선택을 받았는데 저가와 개장가의 선택 빈도가 높아지면 고가의 선택 빈도가 낮아지는 현상이 발견된다. 반면 종가, 거래량, 종가 변동률은 일관된 패턴을 보인다.

입력 변수의 개수가 4개인 경우 R^2를 기준으로 93.89에서 98.54 사이의 예측 성능을 보이며, 모든 평가 기준을 통해 각 암호화폐에 대해 입력 변수를 1개에서 3개 사용한 경우보다 더 향상된 성능을 보임을 알 수 있다. 4개 변수 조합에서는 고가와 종가가 6개로 가장 높은 선택을 받았고 그 뒤를 이어 거래량과 개장가가 5개의 선택을 받았다. 저가의 선택빈도가 줄어드니 고가의 선택빈도가 높아지는 현상이 발견되었다. 앞선 현상들을 고려했을 때 일정 부분 상호 배타적인 관계가 확인된다.

입력 변수의 개수가 5개인 경우 R^2를 기준으로 95.11에서 99.09 사이의 예측 성능을 보이며, 모든 평가 기준을 통해 모든 암호화폐에서 가장 우수한 예측 성능을 보인다. 또한, 모든 암호화폐에서 종가 변동률을 제외한 모든 변수가 선택되었다. 반면 입력 변수의 개수가 6개에 대한 평가는 90.76에서 97.15 사이의 예측 성능을 보였다. 이것으로 종가 변동률은 성능향상보다는 성능 악화 쪽에 영향을 준다는 것이 보다 명확해졌다.

본 실험을 통해 내린 결론은 다음과 같다. 첫 번째로 통계기반 분석에서 종가와 상관관계가 거의 없어 부정적 영향을 끼칠 수 있다고 검증된 종가 변동률의 경우 딥 러닝 기반 암호화폐 종가 예측모델에 대해서도 동일하게 부정적 영향을 끼침을 모든 평가 기준을 통해 검증했다. 두 번째로 통계기반 분석에서 상관관계가 거의 없으며, 다중 공선성이 존재하지 않아 긍정적 영향을 끼칠 수 있다고 검증된 거래량의 경우 암호화폐 종가 예측에 사용된 입력 변수의 개수가 많아질수록 사용 빈도 또한 일관되게 증가하였으며, 딥 러닝 기반 암호화폐 종가 예측모델에 대해서도 동일하게 긍정적 영향을 끼침을 모든 평가 기준을 통해 검증했다. 세 번째로 통계기반 분석에서 다중 공선성으로 인해 과대 적합을 발생시킬 가능성이 있는 개장가, 고가, 저가의 경우 이 변수들이 모두 사용되는 것이 성능향상에 효과는 있으나 제한된 횟수에서는 일정 부분 상호 배타적인 관계를 보인다. 특히 고가와 저가의 경우 둘 중 하나만 선택되는 경우가 많았다. 이것은 고가이든 저가이든 하나만 선택되면 다른 하나는 꼭 필요한 변수가 아니며 포함되어도 성능향상 폭이 높은 것은 아니다.

5. 결론 및 향후 연구

본 논문에서는 입력 변수들이 암호화폐 종가 예측에 미치는 영향력을 상관분석과 다중 공선성 분석을 기반으로 한 통계기반 분석과 R^2, MSE, RMSE, MAE를 평가 기준으로 사용한 딥 러닝 기반 분석을 통해 입력 변수들이 암호화폐 종가 예측에 미치는 영향력을 분석했다.

통계기반 분석 실험에서는 거래량은 종가형성에 전반적으로 적지 않은 영향력을 미침을 결론 내렸다. 종가 변동률은 다중 공선성이 존재하지 않지만, 종가에 대해 상관관계가 거의 존재하지 않아 입력 변수로 적합하지 않다는 결론을 내렸다. 마지막으로 개장가, 고가, 저가는 종가에 대해 완벽한 상관관계를 가지고 있지만, 다중 공선성이 존재해 독립적인 변수가 아니며, 암호화폐 종가 예측에 과대 적합을 발생시킬 가능성이 있다는 것을 실험 및 검증했다.

딥 러닝 기반 분석 실험에서는 통계기반 분석과 동일하게 거래량은 암호화폐 종가 예측에 긍정적 영향을 끼치며, 종가 변동률은 암호화폐 종가 예측에 부정적 영향을 끼친다는 것을 딥 러닝 기반 암호화폐 종가 예측모델 평가 기준을 통해 실험 및 검증했다. 반면 개장가, 고가, 저가는 통계기반 분석 실험에서는 암호화폐 종가 예측에 과대 적합을 발생시킬 가능성이 있다는 결론을 내렸으나, 딥 러닝 기반 분석 실험에서는 이 변수들이 모두 사용되는 것이 성능향상에 효과는 있으나 제한된 횟수에서는 일정 부분 상호 배타적인 관계를 보인다는 것을 실험 및 검증했다. 특히 고가와 저가의 경우 둘 중 하나만 선택되는 경우가 많았다. 이를 통해 고가이든 저가이든 하나만 선택되면 다른 하나는 꼭 필요한 변수가 아니며 포함되어도 성능향상 폭이 높은 것은 아님을 알 수 있다.

두 실험 결과를 바탕으로 실험에 사용된 입력 변수에 대해 종가 변동률은 예측 성능에 부정적 영향을, 거래량은 예측 성능에 긍정적 영향을 끼침을 결론 내렸다. 또한, 다중 공선성으로 인해 과대 적합을 발생시킬 가능성이 있는 개장가, 고가, 저가의 경우 이 변수들이 모두 사용되는 것이 성능향상에 효과는 있으나 제한된 횟수에서는 일정 부분 상호 배타적인 관계를 보임을 결론 내렸다.

최종적으로 종가 변동률을 제외한 개장가, 고가, 저가, 거래량, 종가를 사용한 조합이 가장 최적의 성능을 보이는 조합임을 결론 내렸다.

본 연구는 20가지 암호화폐 거래소에서 제공하는 모든 입력 변수인 개장가, 고가, 저가, 거래량, 종가 변동률, 종가에 대해서만 통계기반과 딥 러닝 기반 분석을 통해 영향력을 분석하는 실험을 수행했다. 그러나 관련 연구에서 볼 수 있듯이 현재 암호화폐 종가 예측 연구에서는 다양한 입력 변수를 사용한 선행 연구가 있으며, 이를 기반으로 한 연구들이 진행되고 있다. 따라서 유사한 특성들을 지닌 입력 변수들을 카테고리화한 뒤 동일한 실험을 통해 카테고리 내부에서 입력 변수에서 암호화폐 예측 성능에서 입력 변수가 끼치는 영향력을 분석하거나 각 입력 변수 카테고리가 끼치는 영향력을 분석 및 검증하는 연구를 수행할 예정이다. 또한, 이를 기반으로 더 향상된 예측 성능을 보이는 딥 러닝 기반 예측모델에 관한 연구도 함께 수행할 예정이다.

References

[1] H. Zexin, Z. Yiqi, and K. Matloob, "A survey of forex and stock price prediction using deep learning," in *Applied System Innovation*, Vol.4, No.1, pp.1-9, 2021.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, Vol.9, No.8, pp.1735-1780, 1997.

[3] E. Hoseinzade and S. Haratizadeh, "CNNpred: CNN-based stock market prediction using a diverse set of variables," *Expert Systems with Applications*, Vol.129, pp.273-285, 2019.

[4] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Computing and Applications*, Vol.32, No.13, pp.9713-9729, 2020.

[5] H. Yamamoto, H. Sakaji, and H. Matsushima, "Forecasting crypto-asset price using influencer tweets," in *International Conference on Advanced Information Networking and Applications,* pp.940-951, 2019.

[6] H. Maqsood, I. Mehmood, and M. Maqsood, "A local and global event sentiment based efficient stock exchange forecasting using deep learning," *International Journal of Information Management,* Vol.50, pp.432-451, 2020.

[7] Z. Hu, W. Liu, and J. Bian, "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction," in *Proceedings of the eleventh ACM International Conference on Web Search and Data Mining,* pp.261-269, 2018.

[8] Y. Xuan, Y. Yu, and K. Wu, "Prediction of short-term stock prices based on EMD-LSTM-CSI neural network method," in *2020 5th IEEE International Conference on Big Data Analytics,* pp.135-139, 2020.

[9] N. E. Huang, Z. Shen, and S. R. Long, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proceedings of the Royal Society of London. Series A: Mathematical, physical and engineering sciences,* Vol.454, No.1971, pp.903-995, 1998.

[10] R. Hadi and F. Hamidreza, "Stock price prediction using deep learning and frequency decomposition," *Expert Systems with Applications,* Vol.169, pp.1-29, 2021.

[11] Y. Li, P. Ni, and V. Chang, "Application of deep reinforcement learning in stock trading strategies and stock forecasting," *Computing*, Vol.102, No.6, pp.1305-1322, 2020.

[12] B. S. Lin. W. T. Chu, and C. M. Wang, "Application of stock analysis using deep learning," in *2018 7th International Congress on Advanced Applied Informatics*, pp.612-617, 2018.

[13] P. Oncharoen and P. Vateekul, "Deep learning using risk-reward function for stock market prediction," in *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence,* pp.556-561, 2018.

[14] Z. Li, D. Yang, and L. Zhao, "Individualized indicator for all: Stock-wise technical indicator optimization with stock embedding," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* pp.894-902, 2019.

[15] S. Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System [Internet], https://www.ussc.gov/sites/default/files/ pdf/training/annual-national-training-seminar/2018/Emerging_Tech_Bitcoin_Crypto.pdf, Jul. 12, 2021.

[16] V. Buterin, Ethereum Whitepaper [Internet], https://http://kryptosvet.eu/wp-content/uploads/2021/05/ethereum-whitepaper-kryptosvet.eu_.pdf, Jul. 12, 2021.

[17] M. Watorek, S. Drozdz, and J. Kwaipien, "Multiscale characteristics of the emerging global cryptocurrency market," in *Physics Reports,* 2020.

[18] C. Worley and A. Skjellum, "Blockchain tradeoffs and challenges for current and emerging applications: Generalization, fragmentation, sidechains, and scalability," in *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData),* pp.1582- 1587, 2018.

[19] F. Colon, C. Kim, and W. Kim, "The effect of political and economic uncertainty on the cryptocurrency market," *Finance Research Letters,* Vol.39, pp.1-7, 2021.

[20] A. K. Tanwar, S. Kumar, and R. Patthak, "Modelling the dynamics of Bitcoin and Litecoin: GARCH versus stochastic volatility models," in *Applied Economics,* Vol.51, No.37, pp.4073-4082, 2019.

[21] Top Cryptocurrency Spot Exchanges [Internet], https://coinmarketcap.com/rankings/exchanges, July 15, 2021.

[22] M. Patel and S. Tanwar, "A deep learning-based cryptocurrency price prediction scheme for financial institutions," *Journal of Information Security and Applications*, Vol.55, pp.1-13, 2020.

[23] B. Jacob, C. Jingdong, and H, Yiteng, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, pp.1-4, 2009.

[24] T. A. Craney, J. G. Surles, and SR. Long, "Model-dependent variance inflation factor cutoff values," in *Quality Engineering,* Vol.14, No.3, pp.391-403, 2002.

[25] T. Phaladisailoed and T. Numnonda, "Machine learning models comparison for bitcoin price prediction," in *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Bali, Indonesia, pp.506-511, 2018.

박 재 현

https://orcid.org/0000-0002-7807-4048
e-mail : kat3160@yu.ac.kr
2021년 영남대학교 컴퓨터공학과(학사)
2021년 ~ 현 재 영남대학교 컴퓨터공학과 석사과정
관심분야 : Data Mining, Software Engineering, AI, and Deep Learning

서 영 석

https://orcid.org/0000-0002-5319-7674
e-mail : ysseo@yu.ac.kr
2006년 숭실대학교 컴퓨터학부(학사)
2008년 KAIST 전산학과(석사)
2012년 KAIST 전산학과(박사)
2014년 ~ 2016년 한국산업기술시험원(KTL) 선임연구원
2016년 ~ 현 재 영남대학교 컴퓨터공학과 교수
관심분야 : Software Engineering, Artificial Intelligence, Internet of Things, and Big data analysis