

Predictions of Cryptocurrency Prices Based on Inherent Interrelationships

Zhenyuan Wu^{1,*}

¹ Chuyuzhentu Information Technology Center, Hunan, China

*Corresponding author. Email: zywu999@163.com

ABSTRACT

The price of cryptocurrencies is predicted in this paper based on their intrinsic interrelationship with Bitcoin. The Kaggle dataset is gathered, standardized, collated, and extracted. Convolutional Neural Network (CNN) is compared to other machine learning methods such as Linear Regression and K-Nearest Neighbor (KNN), and then parameter optimization is performed. The empirical results show that Linear Regression is less accurate than the other two models, whereas the CNN model employing end-to-end solutions outperforms other models with the best accuracy (overall above 0.95) forecasting the price quantitatively and directly of the majority of cryptocurrencies, despite the fact that forecasting takes a long time and tweaking its parameters is extremely time-consuming. This paper proposes using research object interrelationships rather than extrinsic relationships.

Keywords: *cryptocurrency price prediction, interrelationships, machine learning.*

1. INTRODUCTION

Cryptocurrency is a subset of digital currencies that employs cryptography for protection [1]. The cryptocurrency was initially offered to the public in 2009. While Bitcoin and other cryptocurrencies took years to gain popularity, they have then grown swiftly [2]. In 2008, Bitcoin was valued at \$1. It was about \$27 in late 2009 and then around \$7500 [3]. (Figure 1 evinces the price tendency of Bitcoin.) After Bitcoin, Nakamoto's followers created a lot. The value of Bitcoin, Ethereum, and other cryptocurrencies skyrocketed. Cryptocurrencies like Bitcoin are gaining acceptance as genuine forms of payment as their real-world value and utility are confirmed. Their use cases can be summarized as follows: digital asset/investment speculation, medium of exchange, payment system, and non-monetary use cases [4]. Coins with superior speed and privacy to Bitcoin are gaining traction. There are approximately 1000 cryptocurrencies in existence, while new ones appear on a frequent basis [3]. The study of the crypto market is a young and immature concept [5].-Thus, it is fair to believe that cryptocurrency prediction is vital and necessary.

Cryptocurrency prediction has remained in the spotlight in recent years, analyzed by a wide range of solutions. Jay et al. propose a stochastic neural network model, which is based on the random walk theory and is

a new method of modeling stock prices. The proposed model adds layer-wise randomness to neural network feature activations to simulate market volatility [6]. Patel et al. use a hybrid cryptocurrency prediction scheme based on Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) [7]. Traditional supervised learning algorithms, Logistic Regression, Linear Support Vector Machine, Bernoulli Naive Bayes, were applied by Lamon et al. [8]. In the study by Rebane and colleagues, the performance of ARIMA is compared to that of a seq2seq recurrent deep multi-layer neural network (seq2seq) that involves a series of input types [9]. Tanwar et al. raised a hybrid model based on deep learning (which integrates GRU and LSTM) for predicting the price of Litecoin and Zcash with respect to the parent currency's interdependence [10]. Among all the researches, they seem to concentrate on external links of the cryptocurrencies and ignore their inner relationship. Though Tanwar et al. noticed interrelationships, they still neglect the importance of Bitcoin.

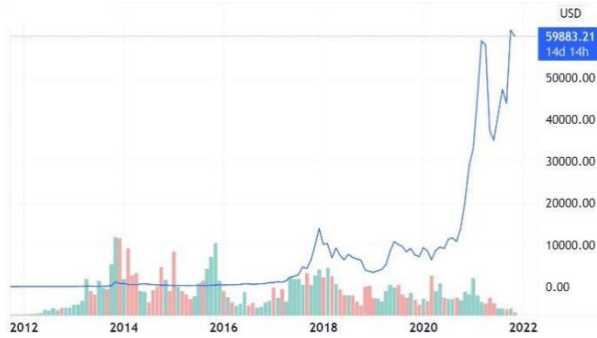


Figure 1. Bitcoin price tendency. *Notes:* This figure is collected from <https://www.tradingview.com/chart/>.

The co-integrated crypto market's strength of dependence among cryptocurrencies has grown in recent years, which may aid investors in risk management while identifying opportunities for alternative diversified and profitable investments [11]. Crypto-coins have been shown to be highly linear and non-linearly correlated [12]. Since 2015, all crypto-coins have been positively related, with the strength of dependence increasing [12]. Bitcoin is linked to all other crypto-currencies. Although the crypto-market is still in its infancy, it is rapidly expanding, with leading crypto-currencies already showing numbers that are comparable to those of real coins/assets traded in mature markets [12].

To predict the price of the top seven cryptocurrencies, with the exception of Bitcoin, information from the original dataset is collected and preprocessed, and then a deep learning method known as the CNN is implemented to predict the price of Ripple, Ether, Stellar, Bitcoin-cash, EOS, and Litecoin one-day later. In order to evaluate the accuracy of the CNN method, it is then compared to other machine learning methods, such as linear regression [13] and KNN [14], which are both widely used in the forecasting domain. Through collation and refinement, the preprocessed data is then processed by models. The result shows that Linear Regression is less accurate than the other two models while CNN is capable of accurately (overall accuracy over 0.95) forecasting the price of the majority of cryptocurrencies, whereas it takes a long to forecast and tweaking its parameters is incredibly time intensive. This article is theoretically innovative in that it predicts using research object interrelationships rather than extrinsic relationships. Not only is it the first of its kind in terms of price forecasting entry points, but it also marks a further refinement in the comprehensiveness of the approach to digital currency price forecasting.

The remainder of this paper is organized in the following manner. Section 2 presents a description and visualization of data, as well as statistical summaries of the information. Section 3 introduces the CNN, Linear regression, and KNN algorithms, as well as some model settings. Parameter optimization and CNN model testing are included in Section 4, which is followed by a comparison of model accuracy, while section 5 draws conclusions.

2. DATA AND VARIABLES

2.1. Data Description

The studied subject of this experiment is cryptocurrencies, whose data are provided by Kaggle and UCI Machine Learning. The dataset includes historical open, high, low, close, trading volume, and market ranking data for all cryptocurrencies. The data was collected between April 28th, 2013, and November 21st, 2018. The collected data provides 942297 observations of 1584 cryptocurrencies, and 12 characteristics are obtained after eliminating missing data.

2.2. Data Visualization and Statistical Summaries

In this project, only the top seven currencies in terms of market capitalization (Bitcoin, Ripple, Ether, Stellar, Bitcoin-cash, EOS, and Litecoin) are analyzed. Table 1 demonstrates details of variables used in this experiment.

Table 1. Variables of cryptocurrency's dataset.

	Type	Description
Slug	non-numerical	name of the cryptocurrency
Symbol	non-numerical	abbreviation of the cryptocurrency
Name	non-numerical	name of the cryptocurrency
Date	non-numerical	₩
Rank now	numerical	ranking of the cryptocurrency's market cap
Open	numerical	open price of the cryptocurrency selected
High	numerical	cryptocurrency's highest price selected
Low	numerical	cryptocurrency's lowest price
Close	numerical	close price of the cryptocurrency
Volume	numerical	the trading volume of the day
Market	numerical	market cap of the day
Close_ratio	numerical	Close Ratio = (Close-Low)/(High-Low)

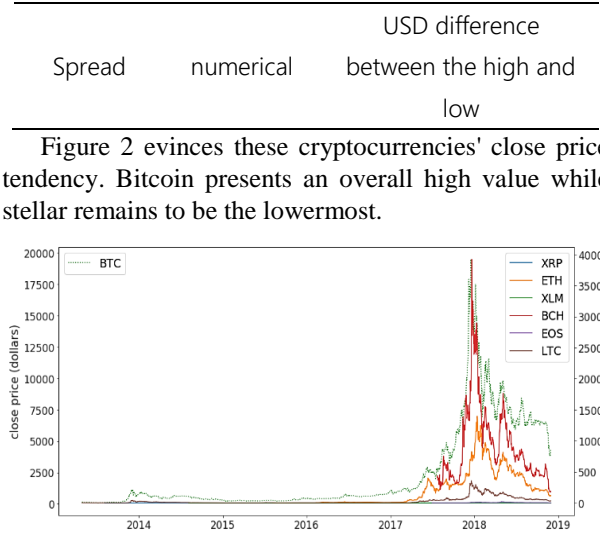


Figure 2. Close price for the top seven cryptocurrencies.

3. METHODOLOGY

3.1. Convolutional Neural Network (CNN)

CNN's are specialized neural networks for processing input with a specified grid-like architecture. This includes time-series data, which is 1D, and image data, which is a 2D grid of pixels [15]. It's a special case of linear. In one or more of their layers, convolutional networks use convolution instead of matrix multiplication. To identify long-term dependencies in given data, these three models are used [15]. The previous layer feature maps are convolved with learnable kernels, which, through the activation function, form the output feature map. With convolutions, multiple input maps can be combined as the output [16]. The convolutional layer is the brain of a CNN, and it is responsible for producing feature maps by applying trained filters to input data. Applying and sliding a one-dimensional (time) filter over a time sequence is what convolution is [17].

The process of the CNN model is shown in Figure3. a^l refers to the output of the hidden layer while a^{l-1} stands for inputs of the hidden layer, σ means activation function, and σ is offset values.

$$a^l = \sigma z^l = \sigma(a^{l-1} \times W^l + b^l) \quad (1)$$

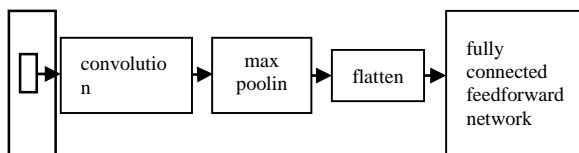


Figure 3. Convolutional Neural Network (CNN) flow path.

Pooling layers were proposed to alleviate the issue that convolutional feature maps record the exact position of features in the input, while some employ batch normalization and dropout layers to speed up training and reduce overfitting.

3.2. K-Nearest Neighbor (KNN)

KNN is a simple but powerful example-based text categorization method. This algorithm's notion is as follows: The algorithm searches for the k nearest neighbors among the pre-classified training documents based on some similarity measure and ranks those k neighbors based on their similarity scores; the categories of the KNNs are used to predict the category of the test document by using the ranked scores of each as the weight of the candidate categories, if more than one neighbor belongs to the same category, the sum of their scores is used to determine the weight of that category; if the test document exceeds a predetermined threshold, the category with the highest score is assigned to it; and the test document can be assigned to multiple categories [18].

When considering the similarity between the unknown sample and the known sample, L_p is introduced. L_p Refers to distance. $x_i \in R^n, x_j \in R^n$

$$L_p(x_i, x_j) = (\sum_{l=1}^n |x_i^{(l)} - x_j^{(l)}|^p)^{\frac{1}{p}} \quad (2)$$

)

Manhattan Distance is applied when $p = 1$.

$$L_1 = \sum_{k=1}^n |x_{1k} - x_{2k}| \quad (3)$$

)

Euclidean Distance is utilized when $p = 2$.

$$d_{12} = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2} \quad (4)$$

)

When $p \rightarrow \infty$, Chebyshev Distance is put to use.

$$d_{12} = \max(|x_1 - x_2|, |y_1 - y_2|) \quad (5)$$

)

When features of the data are not on the same scale, one way is to replace x_{im} by z_{im} .

$$z_{im} = \frac{x_{im} - \bar{x}_m}{\sigma_m} \quad (6)$$

)

$$\bar{x}_m = \frac{1}{N} \sum_{i=1}^N x_{im} \quad (7)$$

)

$$\sigma_m^2 = \frac{1}{N} \sum_{i=1}^N (x_{im} - \bar{x}_m)^2 \quad (8)$$

)

\bar{x}_m refers to the empirical mean of m th feature, σ_m^2 refers to the empirical variance of m th feature.

3.3. Linear Regression

Linear regression represents the relationship between numerous relevant predictors and a particular variable. In other words, a linear model links the explanatory predictor dataset components to the resultant variables. The least-squares method is widely used in linear regression [19]. Regression is a statistical method that allows actions on a dataset with target values. Providing more data can also improve the outcome. Regression creates patterns by correlating predictor and target values. This pattern applies to many datasets with ambiguous target values. This divides the data into two sections: model definition and model testing [20]. We started by categorizing the data into training and testing. The training phase is then utilized to start analyzing and defining the model.

In general, regressor variables can be linked to the response variable y . A multiple linear regression model with p regressor variables is the following model. The regressor variables x_i is described by a hyperplane in p -dimensional space in this model. In a multiple linear regression model, the least-square approach is commonly used to estimate the regression coefficients (w_j) while intercept is designated to be w_0 [20].

$$\hat{y}(w, x) = w_0 + w_1x_1 + \dots + w_px_p \quad (10)$$

4. EMPIRICAL RESULTS

4.1. Data Preprocessing

To evaluate correlations between Bitcoin and the other top-7 cryptocurrencies, lists are extracted from the original dataset. By deleting redundant columns and blanks, I was able to acquire useful data, which were then adjusted to reduce magnitude effects [21]. For analytical purposes, formula (11) is used to normalize the data. Through this function, data were mapped to the range $[-1, 1]$. Normalized data visualization is presented in Figure 4. As a consequence, each cryptocurrency's normalized close price can be compared. I refers to original data while I_{mean} represents the mean value of the raw data. I_{max} and I_{min} are the maximum value and the minimum value of the original data while I_{new} are the normalized data.

$$I_{new} = \frac{I - I_{mean}}{I_{max} - I_{min}} \quad (11)$$

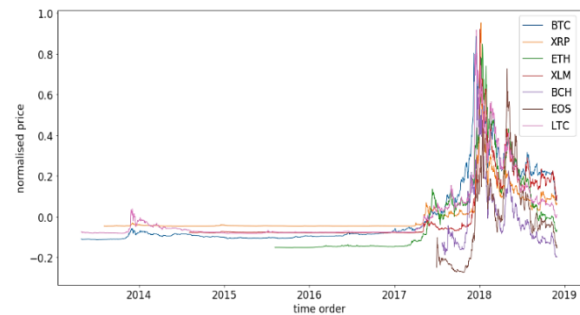


Figure 4. Close price (Normalized) comparison between Bitcoin and other cryptocurrencies.

According to Figure 5, correlation coefficients of the coin's next-day close price and the close price of Bitcoin are obvious. Predicted Litecoin price is strongly correlated with Bitcoin while Ether evinces a negative relationship. The correlation coefficients were calculated with the Pearson method.

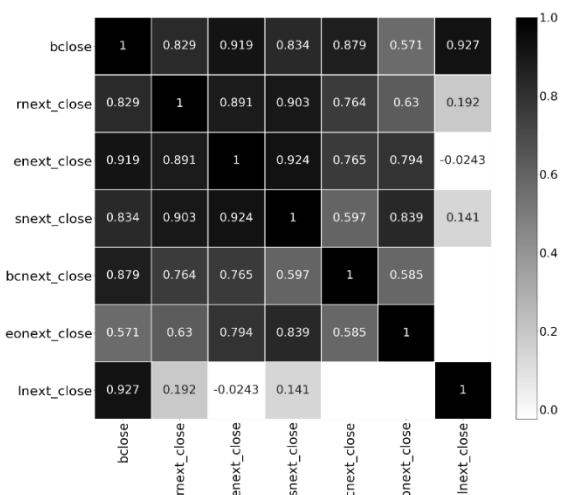


Figure 5. heatmap of predicted prices and bitcoin close price.

4.2. Model Selection

In this section, the accuracy of CNN, Linear Regression, and K-nearest neighbor algorithms will be evaluated, and the algorithm that performs the best will be chosen as the experiment's model. Training (80%) and testing (20%) were repeated to confirm experiment stability, and the training and testing sets were split from the preprocessed dataset. Because of their high prediction accuracy, CNN, linear regression, and KNN were chosen to be trained. To evaluate the models, a score from sklearn is introduced, and Table 3 illustrates the results. For Linear Regression and KNN, the square of multiple correlation coefficient R^2 is returned [22]. When $R^2 > 0.4$, the linear trend appears considerable, and linear regression can be used safely, while for CNN, I got mean accuracy [22].

As shown in Table 2, the CNN model has a comparable high accuracy when forecasting the prices of XRP, XLM, BCH, EOS, and LTC, even reaching 0.980

when forecasting the Stellar price. KNN has the second-best overall score, while linear regression has the lowest score, even reaching 0.326 when forecasting the price of EOS. Thus, CNN's maybe the most capable of forecasting the direction and magnitude of XRP, ETH, XLM, BCH, and LTC's closing prices. Linear regression looks to be the weakest model in this trial.

Table 2. The correlation coefficient of cryptocurrencies' predictions and Bitcoin price.

	Correlation coefficient
XRP next day-BTC	0.83
ETH next day -BTC	0.82
XLM next day -BTC	0.83
BCH next day -BTC	0.88
EOS next day -BTC	0.57
LTC next day -BTC	0.93

Table 3. Accuracy comparison of cryptocurrencies' price prediction using Bitcoin.

	CNN (Mean accuracy)	Linear Regression (R^2)	KNN (R^2)
XRP next day -BTC	0.963	0.687	0.956
ETH next day -BTC	0.921	0.845	0.972
XLM next day -BTC	0.980	0.696	0.973
BCH next day -BTC	0.970	0.772	0.804
EOS next day -BTC	0.972	0.326	0.967
LTC next day -BTC	0.951	0.859	0.936

The parameter optimization is proved potent for infinity as a method of model accuracy enhancement. For CNN, the original state appeared to be a shallow design with only one hidden layer and 100 hidden neurons. With a learning rate of $\eta = 0.1$, a mini-batch size of 10, and no regularization, we'll train for 60 epochs. Then stride lengths of 1 and five by five local receptive fields are used. We'll also add a max-pooling layer that combines two by two pooling windows [23]. Thus, the accuracy of the model is improved. As for Stellar, the prediction's accuracy increased from 0.980 to 0.9898. Rectified Linear Unit (ReLU) nonlinearity is also applicable, as it allows the convergence of extremely deep models training [24].

4.4. Model Testing

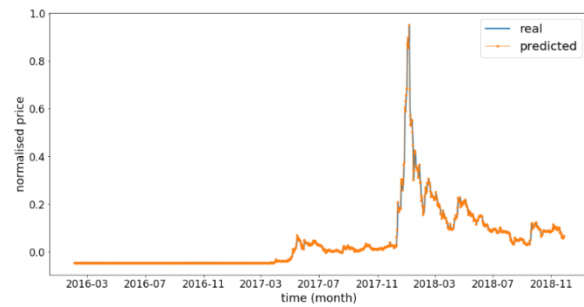


Figure 6. Actual price and CNN-predicted price comparison (XRP).

Even in locations where the price varies violently, the lines indicating the true price and the CNN's prediction price of each cryptocurrency are tightly aligned. The linear regression's prediction accuracy is adequate in the stable area, but it is inferior in the area with dramatic variations. Linear regression's potential to predict price fluctuations is limited, which hinders the model's accuracy. As a result, CNN is able to precisely predict the price of most cryptocurrencies as well as their variations. According to Table 3, except for slightly worse performance when predicting ETH, the model's accuracy in predicting currency prices was greater than 95%, significantly higher than the accuracy of the other two models in the experiment. The reason for this could be that CNN can extract meaningful information from features by abstracting many levels of representation, resulting in improved prediction ability. However, it takes significantly longer to predict, and optimizing its parameters is extremely time-consuming.

5. CONCLUSION

The purpose of this paper is to predict cryptocurrency's price based on the interrelationships that exist between Bitcoin and the other top seven cryptocurrencies. After processing the data collected from Kaggle, in order to reach the target, this paper applies machine learning techniques (CNN, Linear Regression, and KNN) and chooses the one that has the highest degree of precision. According to the factors discussed previously, CNN's accuracy surpasses that of the other two popular machine learning algorithms, KNN and linear regression, in both directional and quantitative prediction. There are two main outcomes. Firstly, the feature system constructed using aspects of the cryptocurrency market and the interaction of bitcoin and other cryptocurrencies is effective and provides some additional insights into the factors that contribute to the price fluctuations of XRP, ETH, XLM, BCH, and LTC. Furthermore, the deep learning method is a good way to estimate cryptocurrency prices. Deep learning retrieves high-level features from data directly, eliminating the need to develop a feature extractor for each task, whereas traditional machine learning decomposes a problem into

several sub-problems, which are then solved individually before taking the sum of all sub-problems to produce the final results.

These extrapolations serve as a guide for investors and a reference for regulatory agencies in revising their regulations, as well as evidence for economic studies. Because of the similarities and correlations between digital currencies, the objects studied in this project might be used to represent the cryptocurrency market, allowing the deep learning algorithms to anticipate the prices of other cryptocurrencies.

The study's data sources and analyses are insufficient and limited. Only two sorts of features are utilized for prognosis. Text mining and social network information analysis will be employed to collect data with more features and dimensions in the future to undertake a more comprehensive examination of crypto-price prediction. Additionally, this paper did not include all successful machine learning techniques in our analyses. As to enhance and optimize this research, the further study intends to look at novel methodologies such as Adaboost and Recurrent Neural Network (RNN).

AUTHORS' CONTRIBUTIONS

Z. Wu performed the experiment, contributed to the analysis and wrote the manuscript.

REFERENCES

- [1] L. Gilpin, "10 things you should know about bitcoin and digital currencies", *TechRepublic.*, Jun. 2014.
- [2] R. S. Madey, "A study of the history of cryptocurrency and associated risks and threats," Doctoral dissertation, Utica College, 2017.
- [3] K. K. Thakur and G. G. Banik, "Cryptocurrency: its risks and gains and the way ahead," *IOSR Journal of Economics and Finance*, vol. 9, no. 2, pp. 38-42, 2018.
- [4] M. Rauchs and G. Hileman, *Global Cryptocurrency Benchmarking Study*, Cambridge Centre for Alternative Finance Reports, Cambridge Centre for Alternative Finance, Cambridge Judge Business School, University of Cambridge, no. 201704-gcbs, pp. 26, 2017.
- [5] A. Yu. Mikhailov, "Cryptocurrency Market Development: Hurst Method," *Finance: Theory and Practice*, vol. 24, no. 3, pp. 81-91, 2020.
- [6] P. Jay, V. Kalariya, P. Parmar, S. Tanwar, N. Kumar and M. Alazab, "Stochastic neural networks for cryptocurrency price prediction," *IEEE Access*, vol. 8, 2020, pp. 82804-82818.
- [7] M. M. Patel, S. Tanwar, R. Gupta and N. Kumar, "A deep learning-based cryptocurrency price prediction scheme for financial institutions," *Journal of Information Security and Applications*, vol. 55, 2020, pp. 102583.
- [8] C. Lamon, E. Nielsen and E. Redondo, "Cryptocurrency price prediction using news and social media sentiment," *SMU Data Sci. Rev* vol. 1, no. 3, 2017, pp: 1-22.
- [9] J. Rebane, I. Karlsson, P. Papapetrou and S. Denic, "Seq2Seq RNNs and ARIMA models for cryptocurrency prediction: A comparative study," *SIGKDD Fintech'18*, London, UK, 2018, pp. 19-23.
- [10] S. Tanwar, N. P. Patel, S. N. Patel, J. R. Patel, G. Sharma, and I. E. Davidson, "Deep learning-based cryptocurrency price prediction scheme with inter-dependent relations," *IEEE Access*, 2021.
- [11] T. Stengos, "Recent developments in cryptocurrency markets: co-movements, spillovers and forecasting," *Journal of Risk and Financial Management*, vol. 14, no. 3, 2021, pp. 91.
- [12] B. Vaz de Melo Mendes, and A. Fluminense Carneiro, "A comprehensive statistical analysis of the six major crypto-currencies from August 2015 through June 2020," *Journal of Risk and Financial Management*, vol. 13, no. 9, 2020, pp. 192.
- [13] M. Poongodi, A. Sharma, V. Vijayakumar, V. Bhardwaj, A. P. Sharma, R. Iqbal and R. Kumar, "Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system," *Computers & Electrical Engineering*, vol. 81, 2020, pp. 106527.
- [14] K. Cortez, M. P. Rodríguez-García and S. Mongrut, "Exchange market liquidity prediction with the K-Nearest Neighbor approach: crypto vs. fiat currencies," *Mathematics*, vol. 9, no. 1, 2020, pp. 56.
- [15] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," *IEEE 2017 international conference on advances in computing, communications and informatics (icacci)*, September 2017, pp. 1643-1647.
- [16] S. Liu, C. Zhang and J. Ma, "CNN-LSTM neural network model for quantitative strategy analysis in stock markets," *International conference on neural information processing*, 2017, pp. 198-206.
- [17] I. E. Livieris, E. Pintelas, S. Stavroyiannis and P. Pintelas, "Ensemble deep learning models for forecasting cryptocurrency time-series," *Algorithms*, vol. 13, no. 5, 2020, pp. 121.

- [18] R. Al-Shalabi, G. Kanaan and M. Gharaibeh, "Arabic text categorization using KNN algorithm," *Proceedings of The 4th International Multiconference on Computer Science and Information Technology*, vol. 4, 2006, pp. 5-7.
- [19] X. Huang, L. Gao, R. S. Crosbie, N. Zhang, G. Fu, and R. Doble "Groundwater recharge prediction using linear regression, multi-layer perception network, and deep learning," *Water*, vol. 11, no. 9, 2019, pp. 1879.
- [20] S. N. Ogulata, C. Sahin, R. T. Ogulata and O. Balci, "The prediction of elongation and recovery of woven bi-stretch fabric using artificial neural network and linear regression models," *Fibres & Textiles in Eastern Europe*, vol. 14, no. 2(56), 2006, pp. 46-49.
- [21] S. C. Nayak, B. B. Misra and H. S. Behera, "Impact of data normalization on stock index forecasting," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 6, 2014, pp. 2150-7988.
- [22] W. Hämmäläinen and M. Vinni, "Comparison of machine learning methods for intelligent tutoring systems," *International conference on intelligent tutoring systems*, 2006, pp. 525-534.
- [23] M. Nielsen, *Neural Networks and Deep Learning*, 2019, ch. 6.
- [24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich "Going deeper with convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.