IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⧓ Impact Factor 7.39 ⧓ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116121

CRYPTOCURRENCY PRICE PREDICTION USING MACHINE LEARNING

Nilesh Shrenik Hosure[1], Jagadish V Gaikwad[2], Shravani M R[3], Nikita Kulloli[4], Madhusudhan H S[5]

Student, NIE Institute of Technology Mysore, Karnataka[1-4]

Assistant professor, NIE Institute of Technology, Mysore, Karnataka[5]

Abstract: After the boom and bust of cryptocurrencies' prices in recent years, it has been increasingly regarded as an investment asset. Because of its highly volatile nature, there is a need for good predictions on which to base investment decisions. Although existing studies have leveraged machine learning for more accurate Cryptocurrency price prediction, few have focused on the feasibility of applying different modelling techniques to samples with different data structures and dimensional features. To predict Cryptocurrency price at different frequencies using machine learning techniques, we first download the dataset from a trusted website which keeps all the data of various cryptocurrencies then we classify various Cryptocurrencies by the dataset that is according to the available price. We extract the basic trading features acquired from a cryptocurrency exchange are used for 1 month price prediction. Machine learning algorithms including ARIMA and SVR models for Cryptocurrency's daily price prediction with high-dimensional features achieve an accuracy of 93% and 94% respectively, outperforming more complicated machine learning algorithms. Compared with benchmark results for daily price prediction, we achieve a better performance, with the highest accuracy of the machine learning algorithm of 97%. Our Hybrid Machine learning model including Support Vector Regression and Autoregressive integrated moving average for One month's Cryptocurrency price prediction is superior to other Machine learning methods, with accuracy reaching 97%. Our investigation of Cryptocurrency price prediction can be considered a pilot study of the importance of the sample dimension in machine learning techniques.

Keywords: Cryptocurrency, Autoregressive Integrated Moving Average (ARIMA), Support Vector Regression (SVR), Price Prediction, Bitcoin, Transactions, Accuracy, Machine Learning (ML), Deep Learning (DL).

I. INTRODUCTION

A cryptocurrency, crypto-currency, or crypto is a digital currency designed to work as a medium of exchange through a computer network that is not reliant on any central authority, such as a government or bank, to uphold or maintain it. There are 18000 various cryptocurrencies', the most popular among them is Bitcoin. Bitcoin is a crypto currency used worldwide for digital payment or simply for investment purposes. Bitcoin is decentralized i.e. it is not owned by anyone. Transactions made by bitcoins are easy as they are not tied to any country. Investment can be done through various marketplaces known as "bitcoin exchanges." These allow people to sell/buy bitcoins using different currencies. The largest bitcoin exchange is Mt Gox. Bitcoins are stored in a digital wallet which is basically like a virtual bank account. The record of all the transactions, the timestamp data is stored in a place called Blockchain. Each block contains a pointer to a previous block of data. The data on blockchain is encrypted. During transactions the users name is not revealed, but only their wallet ID is made public.

Many cryptocurrencies entered into the crypto market after Bitcoin, for example, Ethereum, launched in 2015, is the second-largest cryptocurrency with a $410 billion market capitalization. More than 5,600 different cryptocurrencies are traded in around 1,100 exchanges and Ripple, Tether, Cardano, Stellar, Litecoin, and Zcash are the most popular digital currencies. Back in June 2016, the total market capitalization of all cryptocurrencies was approximately 12.22 billion dollars and it fluctuated in 2017. It increased to $1.75 trillion in June 2021 [12], with an all-time high of $2 trillion. It will reach nearly $8 trillion by 2030. The daily volume of the crypto market is around $117 billion and more than 100 million people are using these currencies.

Predicting the future is no easy task. Many have tried and many have failed. But many of us would want to know what will happen next and would go to great lengths to figure that out. Imagine the possibilities of knowing what will happen in the future! Imagine what you would have done back in 2012 when Bitcoin was less than $15 knowing that it would surpass $18,000! Many people may regret not buying Bitcoin back then but how were they supposed to know in the first place? This is the dilemma we now face in regards to cryptocurrency. We do not want to miss out on the next jump in price but we do not know when that will or will not happen. So how can we potentially solve this dilemma? Maybe machine learning can tell us the answer.

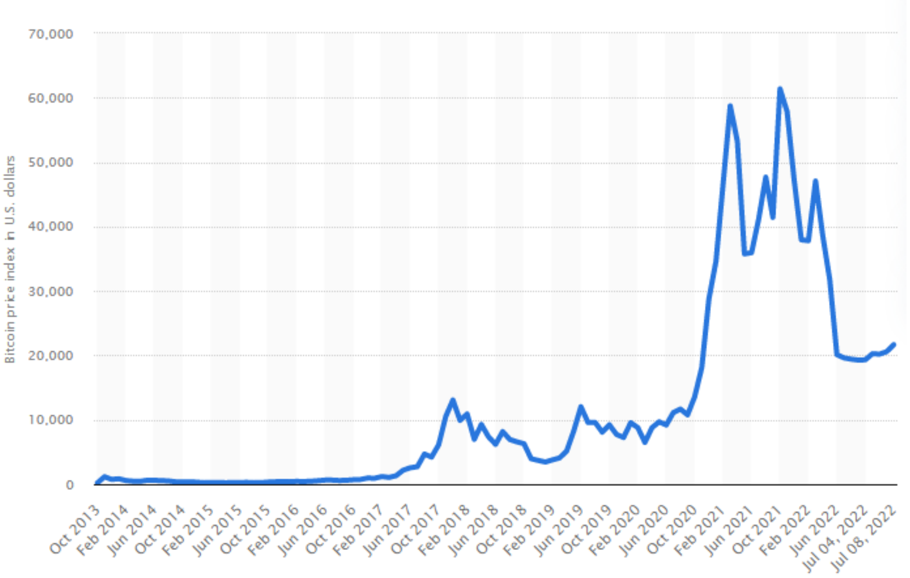


Fig. 1.1 Shows the Bitcoin price variation over the years

Machine learning models can likely give us the insight we need to learn about the future of cryptocurrency. It will not tell us the future but it might tell us the general trend and direction to expect the prices to move. Let's try and use these machine learning models to our advantage and predict the future of Bitcoin by coding them out in Python!

II. RELATED WORK

After the boom of cryptocurrencies' prices in recent years, cryptocurrency has been increasingly regarded as an investment asset. Because of its highly volatile nature, there is a need for good predictions on which to base investment decisions. Although existing studies have leveraged machine learning for more accurate Bitcoin price prediction, few have focused on the feasibility of applying different modelling techniques to samples with different data structures and dimensional features. To predict Bitcoin price at different frequencies using machine learning techniques, many have tried predicting the prices of various cryptocurrencies' using various Machine learning algorithms and some also using mathematical and statistical methods. Most of these predictions resulted in failure or inconsistency in the models. Some of them have also considered various external parameters including twitter sentiment analysis or the effect of covid-19 on the rise and fall of various cryptocurrencies. We have studied specific models which we needed for our predictions, there has also been news about movement of prices of cryptocurrencies based on political mis-beliefs on cryptocurrencies, thus resulting in the fall of prices.

Deep Learning Models: Appropriate design of deep learning models in terms of network parameters is imperative to their success. The three main options available when choosing how to select parameters for deep learning models are random search, grid search and heuristic search methods such as genetic algorithms. Manual grid search and Bayesian optimization were utilized in this study. Grid search, implemented for the Elman RNN, is the process of selecting two hyperparameters with a minimum and maximum for each. One then searches that feature space looking for the best performing parameters. This approach was taken for parameters which were unsuitable for Bayesian optimization. This model was built using Keras in the Python programming language. Similar to the RNN, Bayesian optimization was chosen for selecting LTSM parameters where possible. This is a heuristic search method which works by assuming the function was sampled from a Gaussian process and maintains a posterior distribution for this function as the results of different hyperparameter selections are observed. One can then optimize the expected improvement over the best result to pick hyperparameters for the next experiment. The performance of both the RNN and LSTM network are evaluated on validation data with measures to prevent overfitting. Dropout is implemented in both layers, and we automatically stop model training if its validation loss hasn't improved in 5 epochs.

ARIMA (Autoregressive Integrated Moving Average): The basic principle of the ARIMA model is to estimate the trend and the seasonality of the series and to remove them from the series in order to obtain a stationary series. In this

DOI: 10.17148/IJARCCE.2022.116121

series, statistical forecasting techniques can be used. The final step would be to convert the forecast values to the original scale by applying constraints on trend and seasonality. Trend – mean varying over time. For example, we can see an average increase in the number of bitcoin price over time. Seasonality – time frame variations. For example, people may tend to buy cars in a given month due to pay increments or festivals. Static forecasts are performed and the RMSE is calculated to compare with other models. ARIMA model was implemented to compare its predictability with the LSTM and figure out which is the most suitable method for time series data which has huge fluctuations. ARIMA (Auto regression integrated moving average) is a class model that captures a suite of different standard temporal structures in time series data which include trend, seasonality, cycles, errors and nonstationary data. This allows it to exhibit dynamic temporal behavioural a time sequence. The data preparation phase is done similar to the LSTM model approach [2].

SVM (Support Vector Machine): Predictive models for classification and regression problem is a Support Vector Machine (SVM) works by creating a higher-dimensional model which assigns each new data provided to one category or another. Decision is obtained from the verge that which maximizes the functional and geometric margins between classes. SVM model is best for resolving classification problems [7].

LSTM (Long Short-Term Memory): Long Short-Term Memory Network Model (LSTM) is unique type of Recurrent Neural Network (RNN) specifically used to cease long-term dependency problem. Information is stored for longer period; this marks as their default behaviour. It selects the necessary information to be store and discards the irrelevant information. The sequential characteristics in the time series data is ignored by many of the machine learning algorithms. This problem is identified in rolling window LSTM. Application of any other machine learning classifiers on time series data usually overlook the sequential data. Rolling window LSTM builds model structure for time series prediction by not considering weights. LSTM takes longer time to train as compared to RNN model, the parameters and activation functions increased the computation leading to long term memory [8].

III. PROPOSED METHODOLOGY

Proposed System:

The proposed model has transformed the data in a suitable form which is passed to the machine learning algorithms. Features of the data are extracted and used for prediction. Our hybrid model combines both the models (SVR and ARIMA), we use .merge method to merge both the data frames. Firstly, we collect data from websites (coinmarketcap.com and finance.yahoo.com) then we process the data and split the data and run both the models separately and test them then we apply the .merge to combine both the data frames, it only considers the test values of both the algorithms that are same, other values are deleted in this method. Thus, providing us with the better results of both models.

Methodology:

The system proposed here work on various cryptocurrencies. The method which is applied here is as follows

a. Collect the historical data from any trusted websites (like coinmarketcap.com or finance.yahoo.com).

b. Upload the dataset to our webapp.

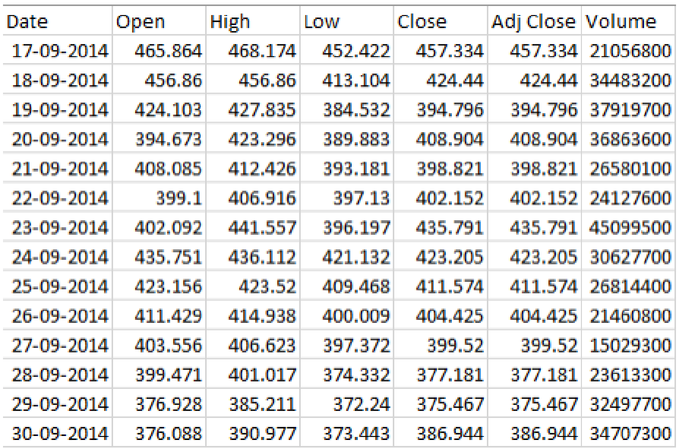
c. The uploaded data is processed.

d. The price of a cryptocurrency after a month is predicted.

e. Also we get the highest and lowest prices during the month.

Data Collection: The dataset requirement entirely depends on the project requirement. The dataset can be collected from various sources such as file, database or even a sensor. The dataset of the various cryptocurrencies' prices used in this work to build machine learning model was collected from finance.yahoo.com and some of the rows and columns of the dataset are shown in fig 3.1. The complete dataset has 7 columns for each cryptocurrency. The CSV file of dataset has prices based upon various factors such as open, high price, Adj. close price, closing market price, market volume.

DOI: 10.17148/IJARCCE.2022.116121



Date	Open	High	Low	Close	Adj Close	Volume
17-09-2014	465.864	468.174	452.422	457.334	457.334	21056800
18-09-2014	456.86	456.86	413.104	424.44	424.44	34483200
19-09-2014	424.103	427.835	384.532	394.796	394.796	37919700
20-09-2014	394.673	423.296	389.883	408.904	408.904	36863600
21-09-2014	408.085	412.426	393.181	398.821	398.821	26580100
22-09-2014	399.1	406.916	397.13	402.152	402.152	24127600
23-09-2014	402.092	441.557	396.197	435.791	435.791	45099500
24-09-2014	435.751	436.112	421.132	423.205	423.205	30627700
25-09-2014	423.156	423.52	409.468	411.574	411.574	26814400
26-09-2014	411.429	414.938	400.009	404.425	404.425	21460800
27-09-2014	403.556	406.623	397.372	399.52	399.52	15029300
28-09-2014	399.471	401.017	374.332	377.181	377.181	23613300
29-09-2014	376.928	385.211	372.24	375.467	375.467	32497700
30-09-2014	376.088	390.977	373.443	386.944	386.944	34707300

Fig 3.1 Sample dataset

Data Pre-processing: This is the most important step in machine learning which helps in building the model more accurately to perform the analysis. In this step, data collected from the site is converted to the clean data set. Then, the data is split into testing set and the training set. For e.g., Data can be divided into 70% of training data while 30% of testing data.

Data Scaling Phase: In this step the data is scaled according to model requirements. It reshapes data to make the model more suitable.

Model Building Phase: The pre-processed data is used to build the best performing model.

Model Learning Phase: After the training data is defined, it is configuring with the defined model to start the learning phase. After the fully configured machine learning algorithms are defined, the data is passed to the model for training. This is achieved by calling fit () method.

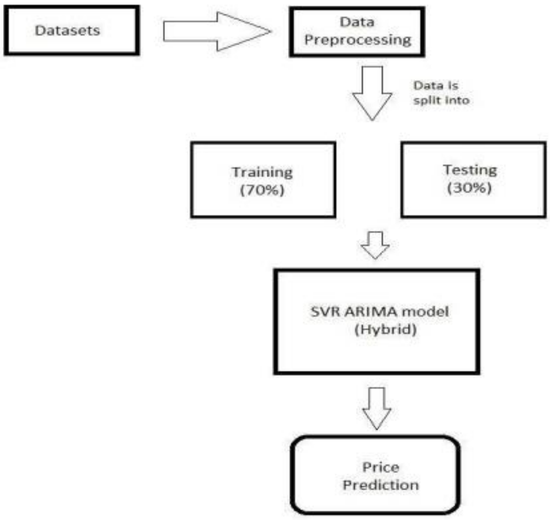


Fig 3.2 Flow chart of our model

Evaluation: This is the integral part as it helps in finding the best model that represents data and how well the prediction is achieved. Input values are passed to the model and the output is the predicted values. The output is compared with the testing data to calculate accuracy and the RMSE values.

IV. SYSTEM DESIGN IMPLEMENTATION

Data collection

Dataset collection is a process of collecting required data from all the relevant sources as needed for the research problem, test the hypothesis and evaluate the outcomes. Yahoo Finance: Yahoo! Finance is a media property that is part of the Yahoo! network. It provides financial news, data and commentary including stock quotes, press releases, financial reports, and original content. It also offers some online tools for personal finance management. In addition to posting partner content from other web sites, it posts original stories by its team of staff journalists. The website has provided us with the historical dataset of various cryptocurrencies.

CoinMarketCap: CoinMarketCap is a cryptocurrency industry utility that aggregates and reports recently-traded prices for hundreds of cryptocurrencies traded on hundreds of platforms around the world. For each currency it reports the total value of the outstanding currency (the "market capitalization"), its total trading volume and its rank by trading volume over the past month and the past 24 hours.

Algorithms used:

ARIMA (Autoregressive Integrated Moving Average): An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. An autoregressive integrated moving average model is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. The model's goal is to predict future securities or financial market moves by examining the differences between values in the series instead of through actual values. An ARIMA model can be understood by outlining each of its components as follows:

a. Autoregression (AR): refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.

b. Integrated (I): represents the differencing of raw observations to allow for the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).

c. Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations. ARIMA Parameters: Each component in ARIMA functions as a parameter with a standard notation. For ARIMA models, a standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

p: The number of lag observations in the model; also known as the lag order.

d: The number of times that the raw observations are differenced; also known as the degree of differencing.

q: the size of the moving average window; also known as the order of the moving average. In a linear regression model, for example, the number and type of terms are included. A 0 value, which can be used as a parameter, would mean that particular component should not be used in the model. This way, the ARIMA model can be constructed to perform the function of an ARMA model, or even simple AR, I, or MA models.

Step 1: The auto-regressive (AR) model expresses the time series xt at time t as a linear regression of the previous p observations, that is:

$$x_t = \alpha + \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t$$

and where εt is the white noise residual term and ϕi are real parameters.

Step 2: The Moving averages (MA) use dependency between residual errors to forecast values in the next period. The model helps to adjust to unpredictable events. The q th order moving average model denoted by MA (q) is defined as follows:

$$x_t = \alpha - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where α and θi are real parameters. The ARMA model combines the power of AR and MA components together. This way, an ARMA (p, q) model incorporates the p th order AR and q th order MA model, respectively.

Step 3: I denote by Φ and θ the AR and MA coefficients vectors. The α and εt captures the intercept and the error term at time t. The complete ARIMA (p, q) model can be seen in detail in equation, that is

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

SVR (Support Vector Regression): Support Vector Regression (SVR) works on similar principles as Support Vector Machine (SVM) classification. One can say that SVR is the adapted form of SVM when the dependent variable is numerical rather than categorical. A major benefit of using SVR is that it is a non-parametric technique. Unlike SLR, whose results depend on Gauss-Markov assumptions, the output model from SVR does not depend on distributions of the

DOI: 10.17148/IJARCCE.2022.116121

underlying dependent and independent variables. Instead, the SVR technique depends on kernel functions. Another advantage of SVR is that it permits for construction of a non-linear model without changing the explanatory variables, helping in better interpretation of the resultant model. The basic idea behind SVR is not to care about the prediction as long as the error (ϵi) is less than certain value. This is known as the principle of maximal margin. This idea of maximal margin allows viewing SVR as a convex optimization problem. The regression can also be penalized using a cost parameter, which becomes handy to avoid over-fit. SVR is a useful technique provides the user with high flexibility in terms of distribution of underlying variables, relationship between independent and dependent variables and the control on the penalty term.

SVR formula

$$\min \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

assume that ξi , ξi* ∗ ≥ 0 and w is the undetermined parameter vector.

One advantage of using the SVR is the ability to apply different kernel methods which they form different formula. Refer below equations, these define the applied kernels method as follow:

$$\text{Linear} = \mathrm{k}\left(x_i, y_j\right) = x_i . y_j$$

where xi, yi are datasets.

$$\text{Polynomial} = \mathrm{k}\left(x_i, y_j\right) = \left(\left(x_i . y_j\right) + p\right)^q$$

Where p and q are the kernel parameters and satisfy the condition p ≥ 0, q ∈ N.

$$\text{Gauss Kernl(RBF)} = \exp(-\|x_i - x_j\|^2/Q$$

where Q ≥ 0.

Hybrid Model: Our hybrid model combines both the models (SVR and ARIMA), we use .merge method to merge both the data frames. Firstly, we run both the models separately and test them then we apply the .merge to combine both the data frames, it only considers the test values of both the algorithms that are same, other values are deleted in this method. Thus, providing us with the better results of both models.

V. TEST CASES

Module	Input	Expectedoutput	Actual output	Result
Data preprocessing	Preprocessedcsv file	Successfully cleansed data	Taking dataset to project	pass
Predictionusing ARIMA	Preprocessedcsv file with reduced feature	Accuracy according to ARIMA	Accuracy according to ARIMA	pass
Prediction using SVR	Preprocessedcsv file with reduced feature	Accuracy according to SVR	Accuracy according to SVR	pass
Classification using Hybrid Model	Preprocessedcsv file with reduced feature	Accuracy and Prediction according to Hybrid Model	Accuracy and Prediction according to Hybrid Model	pass

VI. CONCLUSION WITH FUTURE ENHANCEMENTS.

All in all, predicting a price-related variable is difficult given the multitude of forces impacting the market. Add to that, the fact that prices are by a large extent dependent on future prospects rather than historic data. However, using machine learning algorithms has provided us with a better understanding of cryptocurrencies, and ARIMA architecture. Other features can be considered It is essential to predict the future prices of cryptocurrencies as they will help in making more profits and reducing losses. Thus, our proposed model which gives accurate results, can predict the prices of coins, the highest and lowest prices during that period.

Our system works efficient on the cryptocurrencies' that have larger dataset. The main improvement would be, to predict

DOI: 10.17148/IJARCCE.2022.116121

the newer coins that have launched recently and has less historical dataset

REFERENCES

1. Vaidehi M, Alivia Pandit, Bhaskar Jindal, Minu Kumari and Rupali Singh., “Bitcoin price prediction using machine learning”, In Proceedings of International Journal of Engineering Technologies and Management Research (IJETMR) 2021.
2. S M Raju, Ali Mohammad Tarif, “Real-Time Prediction of Bitcoin Price using Machine Learning Techniques and Public Sentiment Analysis” In Proceedings of International Islamic University, Malaysia.
3. Sean McNally Jason Roche Simon Caton, “Predicting the Price of Bitcoin Using Machine Learning”, In Proceedings of 26th Euromicro International Conference on parallel, Distributed and Network-based Processing, 2018.
4. Karunya Rathan, Somarouthu Venkat Sai, Tubati Sai Manikanta, “Crypto-Currency price prediction using Decision Tree and Regression techniques” In the proceedings of 3rd International Conference on Trends in Electronics and Information(ICOEI), 2019.
5. Yang Li, Zibin Zheng and Hong-Ning Dai,“Enhancing Bitcoin Price Fluctuation Prediction Using Attentive LSTM and Embedding Network”, In the proceedings of School of Data and Computer Science, Sun Yet-sen University, China 2020.
6. Sudhir N. Dhage, Prachi Vivek Rane, ”Systematic Erudition of Bitcoin Price Prediction using Machine Learning Techniques”, In Proceedings of 5th International Conference on Advanced Computing and Communication System(ICACCS) 2019.
7. Sumit Biswas, Mohandas Pawar, Sachin Badole, Nachiket Galande, Sunil Rathod, “Cryptocurrency Price Prediction Using Neural Networks and Deep Learning”, In Proceedings of 7th International Conference on Advanced Computing and Communication System(ICACCS) 2021.
8. Jiayun Luo, ”Bitcoin price prediction in the time of COVID-19”, In Proceedings of Management Science Informatization and Economic Innovation Development Conference(MSIEID) 2020.
9. E Mahendra, H Madan, S Gupta, S V Singh, “Bitcoin Price Prediction Using Deep Learning and Real Time Deployment”, In Proceedings of 7th International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) 2020.
10. Sudeep Tanvar, Nisargh P Patel, Smith N Patel, Gill R Patel, Gulshan Sharma, Innocent E Davidson, “Deep Learning-Based Cryptocurrency Price Prediction Scheme With Inter-Dependent Relations”, 2021
11. Minul Wimalagunaratne, Guhanathan Poravi, “A Predictive Model for the Global Cryptocurrency Market”, ”, In Proceedings of 8th International Conference on Intelligent Systems, Modelling and Simulation 2018.
12. Lekkala Sreekanth Reddy, Dr.P. Sriramya, “A Research On Bitcoin Price Prediction Using Machine Learning Algorithms”,International Journal of Scientific and Technology Research 2020.