South Asian Journal of Engineering and Technology Research Article

Open Access Full Text Article 123(2022) 87-95 DOI: 10.26524/sajet.2022.12.43

Bitcoin Price Prediction Based on Linear Regression and LSTM

Dr. R. Suneetha Rani[1], D. Sri Vinithri Chowdary[2], Ch. Uday Kiran[3], K. Deekshith[4], V. Alekhya[5]

[1]Professor & Head of the Department, Dept. of Information Technology, QIS College of Engineering and Technology, Andhra Pradesh, India

[2,3,4,5] Final Year (B. Tech), Dept. of Information Technology, QIS College of Engineering and Technology, Andhra Pradesh, India

Corresponding author.
Correspondence: Dr. R. Suneetha Rani
E-mail: suneethari@gmail.com

Article info
Received 25 th March 2022 Received in revised form 26 April 2022 Accepted 9 July 2022

Keywords
Bitcoin, Block chain, Crypto currency, Linear Regression, LSTM

Abstract
Forecasting can be used in many fields such as crypto currency prediction, financial entities, supermarkets etc. We get the time series date which we use to feed the data into the algorithm is given by Y finance with this we get refreshed data every day. The stock market prediction or forecasting helps customers and brokers get a brief view of how the market behaves for the coming years. Many models are currently in use Like Regression techniques, Long Short-Term Memory algorithm etc. FB Prophet is proven to perform better than most other Algorithms with better accuracy. From the proposed research and references we have determined Facebook's Prophet algorithm as our forecasting algorithm because it is predicting at better accuracy, low error rate, handles messy data, doesn't bother for null values and better fitting.

1. INTRODUCTION

Bitcoin is a digital crypto currency that operates on an online decentralized network; it can be traded using an online peer-to-peer Bitcoin network that is not reliant on a central bank or a single administrator. Because it is accepted in over 40 countries worldwide (including Germany, Canada, and Croatia), the emergence of new alternative coins has resulted from its growing popularity. Bitcoin is also used to exchange other crypto currencies, products, and services. Since the introduction of this crypto currency in the year 2009, no hacker has been able to infiltrate it due to block chain technology, where each electronic coin is encrypted with a unique digital signature which makes it easier to track and can be trusted. Each owner signs a digital hash from the previous transaction and adding the public key of the next owner before passing it. The price of Bitcoin in January 2017 was 1,000USD and by the end of December 2017, its value went up to16000 USD and its value as on July 2021 is 32818 USD. We can say that the crypto market is very volatile, and among all the crypto currencies in the market, Bitcoin is experienced by most of the investors due to its anonymity and transparency in the system. This research aims to work on the prediction system for Bitcoin using various Machine learning algorithms and deep learning models to predict the price. There are various factors affecting the price of Bitcoin, in this project we will focus on open, close, high, and low factors.

This paper also consists of 6 libraries:

- Pandas: provides an environment to python for creating creative and practical statistical computing for financial data analysis applications.

- Seaborn: effective visualization for betterunderstanding the graphs and charts.

- Scikit-Learn: implementation of various algorithms. It contains a large variety of supervise and unsupervised learning algorithms.

- Tkinter: To create a faster and quicker GUI application, with cooler features including the implementation of CSS support.

- Pickle: serialization and de-serialization of python object structure to store it in a file/database, maintaining program state, and transfer of data over a network.

2. RELATED WORK

The prediction of crypto coins using the SVM and SVM-PSO method is suggested, where they used the day tradingmethod to predict the values of ETH, BTC,XEM, XRP, XLM, LTC. SVM-PSO shows the optimized results. Performance accuracy of different Classifiers differs from coin to coin. However, this paper works only with a machine learning algorithm, and hence the data can be furtherimproved by implementing the Deep Learning concept. The prediction of Bitcoin price using a transaction graph is proposed. The experiment consists of the Baseline, Logistic Regression, SVM, and Neural Network model with an accuracy of 53.4%,54.3%, 53.7 and 55.1%. The feature selection in this paper is based on the Bitcoin block chain network which tends tobe the least informative feature for the prediction of the Bitcoin price. Predicting the crypto currency prices using sentimental analysis and Machine learning concepts like SVM and Random Forest on ETH, BTC, and XRP with BTC being the highest accuracy of 0.72. This accuracy rateis very low since machine learning algorithms were applied and it can be improved by testing with deep learning models. A prediction model was proposed using four major algorithms, GradientBoosted Tree, Neural Network, Ensemble Learning Method (with the best accuracy of92.4%). The prediction system using Log regression, SVM, ANN, and random forestwas proposed and shows that SVM has thebest accuracy regarding a time-scale activity consisting of daily, 15-, 30- and 60-minutes return. Although SVM does tend toshow better results out of the 4 algorithms, the prediction system can still show better results when Deep learning concepts are applied. A linear regression model was usedto predict the various cryptocurrency price using the open, low, and high cost. The experiment shows an accuracy of 99.3%. This paper does consist of a high accuracy rate but the data set used is comparatively small for a model to work on a real-time chart.

3. OBJECTIVE

The objective of this proposed system is todevelop an application which will predict the bitcoin prices in future with decent accuracy. This allows the investors to investwisely in bitcoin trading as the prices of bitcoin have gone up to an exaggerating amount in the last ten years.

Thus, the main objectives of the "Bitcoin Price Prediction" can be stated as follows:

1. Develop an application which will predict the bitcoin prices in future withdecent accuracy.

2. Allow the investors to invest wisely in bitcoin trading as the prices of bitcoin havegone up to an exaggerating amount in the last ten years.

3. Make use of machine learning algorithmsto increase the accuracy of Bitcoin price prediction.

4. LIMITATION

Although crypto trading has become a new trend, the increase in the number of digital coins and the adaptation of block chaintechnology causes the biggest concern i.e., scalability. It is still dwarfed by the number of transactions that, VISA, processes each day. Additional to that is the speed of transaction which the crypto market cannot compete with the players like VISA and MasterCard until the infrastructure delivering these technologies is massively scaled. The crypto market is very volatile and can never be predicted at 100 percent accuracy. The market depends on human sentiment too; you may never know when aperson owning at least 100 Bitcoin can suddenly sell his entire asset and create a big dip in the crypto market. We can neverpredict a human emotion even with theadvanced technology we have in hand. The analysis of any technical chart composes of mainly 3 major topics, the trend and momentum which indicate the direction and strength of direction, support, and resistance which indicates the potential stopping points of those directions, and thepattern in general, which indicates the information about the market psychology. Cryptocurrencies have not been around for long enough to provide sufficientinformation regarding the resistance and key support compared to the stock market, currencies, and commodities. This makes itdifficult to predict and practice.EXISTING SYSTEM

A Cryptologic pioneer, David Chaum, devised Blind Signature technology, which telecommunicates the encoded messages sealing digital signature, and resulted in inventing Ecash. That is the primary commercial cryptocurrency. Bit 29 Coin in 2009 was the new cryptocurrency. cryptocurrencies have been improved with Block Chain based. Ethereum emerged as the developed money which has services and applications in addition to Block Chainsystem in 2015. WEF (World Economic Forum) suggested that the ranking of Blockchain must be the fourth of 12 future technologies in the Global Risks Report. Furthermore, in 10 years, 10 percent of GDP all over the world is expected to be based on Blockchain technology. In April 2019, about 40 major banks around theworld announced that they would experiment CBDC (central bank digital currencies) founded on Blockchain.Blockchain, which is encrypted with trade information on the public or private network, is a diversified ledger shared withrelevant network participants.

Disadvantages:

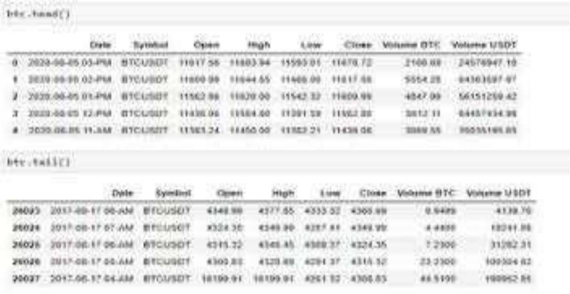
1. In existing system Block Chain Technology is used to predict bitcoin price.
2. By using Blockchain Technology the prices may not be constant they may vary day to day.
3. By using Blockchain Technology we cannot predict the future prices.

6. PROPOSED SYSTEM

The proposed model is used to predict bitcoin price using Machine learning and Neural Network. Machine learning uses Linear Regression and Neural Network uses LSTM for predicting bitcoin prices. In Data Segregation we use features like Open, Close, High, low, Volume BST, Volume UDST, Time, Symbol LinearRegression accuracy rate is 99.87%whereas, LSTM accuracy rate is 97.56%. It is discovered that the Linear Regression model accuracy rate is very high when compared to other models. In this study, wehave used data sets for Bitcoin for testing and training the ML and AI model. With thehelp of python libraries, the data filtration process was done. Python has provided with a best feature for data analysis and visualization. After the understanding ofthe data, we trim the data and use the features or attributes best suited for themodel. Implementation of the model is done and the result is recorded. It was discovered that the linear regression model's accuracy rate is very high when compared to other Machine Learning models from related works; it was found to be 99.87 percent accurate. The LSTM model, on the other hand, shows a mini error rate of 0.08 percent. This, in turn, demonstrates that the neural network modelis more optimized than the machine learning model.

7. METHODOLOGY

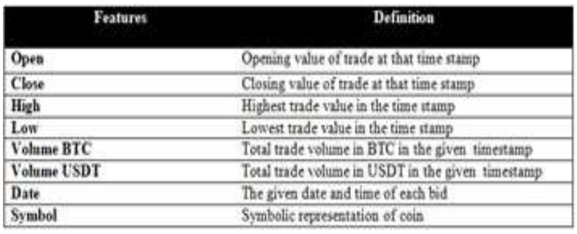
7.1. Data collection: Data Collection is thefirst step we take in order to start any project. It is defined as the procedure of collecting, measuring, and analyzing accurate insights for research usingstandard validation techniques. An analyst would then be able to assess their theory dependent on gathered information. By and large, information assortment is the essential and most significant advance forresearch, independent of the field of examination. The methodology of information assortment is diverse for various fields of study, contingent upon thenecessary data. The most importantobjective of data collection is ensuring thatthe gathered information is rich in content and reliable for statistical analysis so that data-driven decisions can be made efficiently and effectively. The data set contains day transactions from 29th August2017 to 9th August 2020. The data is first tested out with certain regression techniques and then a deep learning model is implemented to provide better accuracy compared to machine learning concepts when there is high or more data sets.



```
btc.head()

                Date   Symbol      Open      High       Low     Close  Volume BTC  Volume USDT
0  2020-08-05 03-PM  BTCUSDT  11617.56  11683.94  11593.01  11678.72     2100.68  24576947.10
1  2020-08-05 02-PM  BTCUSDT  11600.99  11644.55  11466.00  11617.56     5554.28  64363697.97
2  2020-08-05 01-PM  BTCUSDT  11562.98  11620.00  11542.32  11600.99     4847.99  56151259.42
3  2020-08-05 12-PM  BTCUSDT  11438.06  11584.00  11391.59  11562.88     3812.11  44457434.96
4  2020-08-05 11-AM  BTCUSDT  11561.24  11450.00  11352.21  11438.06     3069.55  [illegible]

btc.tail()

                    Date   Symbol      Open      High      Low    Close  Volume BTC  Volume USDT
26023  2017-08-17 08-AM  BTCUSDT   4348.99   4377.85  4333.32  4360.69      0.9499      4139.70
26024  2017-08-17 07-AM  BTCUSDT   4324.35   4349.99  4287.41  4349.99      4.4400     19241.06
26025  2017-08-17 06-AM  BTCUSDT   4315.32   4345.45  4309.37  4324.35      7.2300     31282.31
26026  2017-08-17 05-AM  BTCUSDT   4308.83   4328.69  4291.37  4315.32     23.2300    100304.82
26027  2017-08-17 04-AM  BTCUSDT  16199.91  16199.91  4261.32  4308.83     44.5100    190952.85
```

Fig 1. Display of the Data collected

7.2. Feature Selection: Now that we have the required data for the project, we need tostart the next procedure called data segregation or feature selection. This is a process where we trim out the unwanted data or we remove the unnecessary data from the data set. This step is necessary as we require only those features which can contribute to our prediction as unnecessarydata can cause noise in our final output. To put it in simple words, we segregate data sothat we can have a better model which provides us with an optimized result, reduce the property of over-fitting or redundancy and reduce the training time so that thesystem can generate output faster and with higher accuracy. In this project, I have implemented a few predefined pythonlibraries which help in data visualization and can help you understand the important features which are required by the system. Data visualization is a technique where data or information is represented in a diagrammatic format for better understanding. Data visualization helps us to communicate with the relationships of data using the help of images. These imagesare in form of patterns that can be understood very easily. This is one of the main reasons how machine learning helps in analyzing data. Whether you work in thefinance department or marketing or technical or design, you need to visualize data to understand it. This makes data visualization an important factor in today'sworld.



Features	Definition
Open	Opening value of trade at that time stamp
Close	Closing value of trade at that time stamp
High	Highest trade value in the time stamp
Low	Lowest trade value in the time stamp
Volume BTC	Total trade volume in BTC in the given timestamp
Volume USDT	Total trade volume in USDT in the given timestamp
Date	The given date and time of each bid
Symbol	Symbolic representation of coin

Fig 2. The Features represented in the data

With the help of data visualization libraries, we can see the correlation between features and pinpoint the ones which we require. A sample image is shownbelow to show the correlation graph between the features in the given data set. You can notice in the given image Figure 3

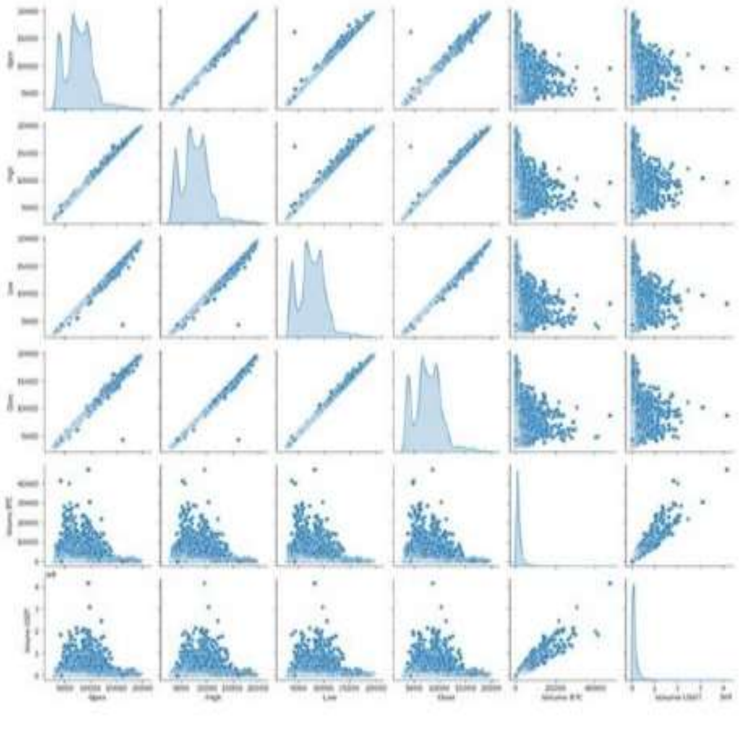


Fig 3. Correlation graph between the features.

7.3. Data Preparation: When variablesthat are measured in different scales it doesnot contribute equally in model fitting which will lead to model learned function to create a bias. Thus, standardization or normalization of data is very much essentialfor better accuracy and result. When working with a Machine Learning model or Deep Learning models where we require back propagation to be more stable andeven faster, proper scaling of data is necessary.

$$x_{scaled} = \frac{x - min\ (x)}{max(x) - min\ (x)}$$

8.ALGORITHMS IMPLEMENTED

8.1. LINEAR REGRESSION

This technique is used to identify the relationship between dependent and independent variables and is leveraged to predict future outcomes. When we use only one dependent and one independent variable then it is called the simple linear regression. As the number of independent and dependent variable increase, it is then referred to as multi-linear regression. The graph is plotted using a straight line across the graph which seeks to be the best fit by calculating the method of least square.

y = mx + C

C = y intercept

m = slope

x, y are the points on the graph

MSE = $\sum 1 \_n \; {}_{i=1}(Y \; {}_{i} \; - \; \hat{Y} \; {}_{i})^{2} \; {}_{n}$

MSE = mean squared errorn = number of data points

Y_i = predicted values

$\hat{Y}$ = predicted values

8.2. LONG SHORT-TERM MEMORY(LSTM)

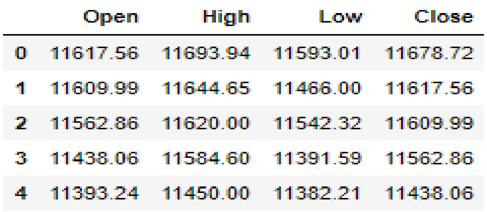
It is a deep learning concept or particularly a Recurrent Neural Networkconcept that avoids the vanishing gradient problem. The main reason for using this algorithm is that it avoids the back propagation error from vanishing or exploding, instead, these errors can flow backward through an unlimited number of virtual layers unfolded in space. LSTM mainly works on time series graphs with data sets that consist of events that occur thousands or millions of discrete-time stepsearlier. It works with given long delaysbetween significant events and can also handle signals with a mixture of low and high-frequency components. Over a lot of researchers have used LSTM to predict time series related data sets for stock prediction and have achieved greater or higher accuracy compared to othery = predictionx = true value

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

y = prediction
x = true value
n = total number of datapoints

9. RESULT AND DISCUSSION

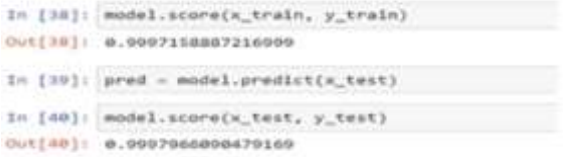
After the data analysis process, we find that the only four features were well suited for the testing of this project. The data was trimmed and only the selected features were left as shown in Figure 4.



	Open	High	Low	Close
0	11617.56	11693.94	11593.01	11678.72
1	11609.99	11644.65	11466.00	11617.56
2	11562.86	11620.00	11542.32	11609.99
3	11438.06	11584.60	11391.59	11562.86
4	11393.24	11450.00	11382.21	11438.06

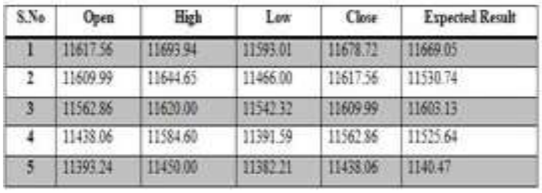
Fig 4.Attribute/Features selected are Open,High, Low, and Close

We can see the output of two models, one which is the Machine Learning model i.e., Linear regression, and the other one is the Recurrent Neural Network model i.e., Long Short-Term Model which shows us the two different outcomes. Linear regression tends to work based on the Mean Squared Equation which tells us the accuracy of the linear graph with respect to the continuous-time frame data set. We see that the accuracy of the training data isapproximately 99.97% and the accuracy ofthe testing data is tending to be approximately 99.97% as shown in Figure5. Meanwhile, the LSTM model tends to find the accuracy with respect to the Mean Absolute Error which shows the error rate approximately to be 0.08% asshown in Figure6.



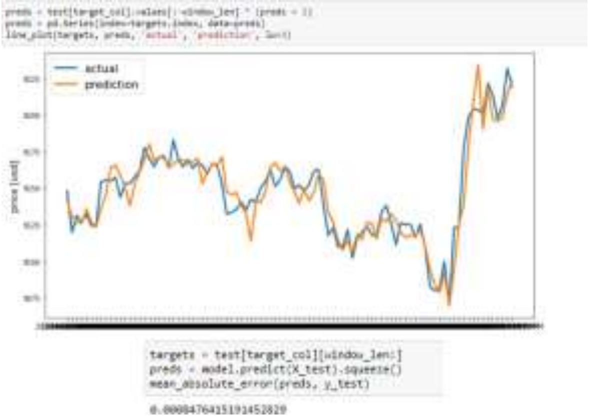
```
In [38]: model.score(x_train, y_train)
Out[38]: 0.9997158887216909

In [39]: pred = model.predict(x_test)

In [40]: model.score(x_test, y_test)
Out[40]: 0.9997966000479169
```

Fig 5. Accuracy obtains from the training andtesting data set using Linear regression model



S.No	Open	High	Low	Close	Expected Result
1	11617.56	11693.94	11593.01	11678.72	11669.05
2	11609.99	11644.65	11466.00	11617.56	11530.74
3	11562.86	11620.00	11542.32	11609.99	11603.13
4	11438.06	11584.60	11391.59	11562.86	11525.64
5	11393.24	11450.00	11382.21	11438.06	1140.47

Fig 6. Testing of Linear regression model



```
preds = test[target_col].values[:-window_len] * (preds + 1)
preds = pd.Series(index=targets.index, data=preds)
line_plot(targets, preds, 'actual', 'prediction', lw=3)
```

```
targets = test[target_col][window_len:]
preds = model.predict(X_test).squeeze()
mean_absolute_error(preds, y_test)

0.0008476415191452829
```

Fig 7. Final Resultant graph of LSTM and theMean Absolute Error rate (0.08%)

DISCUSSION:

The Data visualization shows the correlation between all the features and only the four selected features have a sharp correlation. Data is then fitted into the model using the predefined commands accessible to python. These data models were trained and tested out with a limited number of data sets and provided the result.

With the growing technology and the raise in the data sets we can still work on the model with various other alternative crypto currencies. The model shows a better prediction rate for LSTM but with a very slight difference compared to the linear regression model.

10. CONCLUSION

According to the findings, Long Short-Term Memory has a higher accuracy rate than Linear Regression. Because this study only compares the features of open, close, high, and low, the outcome may alter if we examine additional factors. Data sets cannot be the main rationale for forecasting because the crypto market is dynamic and influenced by social media and otherexternal factors. New data can be acquired,evaluated, and rehearsed as technology progresses, resulting in greater findings forthis experiment.

FUTURE SCOPE

1) More algorithms are being implementedin order to determine the best approach for predicting the crypto currency.

2) Implementing IOT model for smart automatic analysis.

3) To work on a better User Interface so thatpeople can access these data easily and effortlessly.

11. REFERENCES

[1]. Sin E, Wang L. Bitcoin price prediction using ensembles of neural networks. In: 2017 13th International conference on natural computation, fuzzy systems and knowledge discovery. IEEE. 2017;p.666–671. doi:10.1109/FSKD.2017.8393351.

[2]. Shankhdhar A, Singh AK, Naugraiya S, Saini PK. Bitcoin Price Alert and Prediction System using various Models. *IOP Conference Series: Materials Science and Engineering*. 2021;1131(1):012009. Available from:https://dx.doi.org/10.1088/1757-899x/1131/1/012009.

[3]. Mittal R, Arora S, Bhatia MP. Automated cryptocurrencies prices prediction using machine learning. *ICTACTJournal on Soft Computing*.2018;8(4):1758– 1761. Available from: http://ictactjournals.in/paper/IJSC_Vol_8_ Iss_4_Paper_8_1758_1761.pdf.

[4]. Nakamoto S. Bitcoin: A peer-to- peer electronic cash system. 2019. Available from: https://git.dhimmel.com/bitcoin-whitepaper/.

[5]. Sebastião H, Godinho P. Forecasting and trading cryptocurrencieswith machine learning under changing market conditions. *Financial Innovation*. 2021;7(1):1–30. Available from:https://dx.doi.org/10.1186/s40854-020- 00217-x.

[6]. Phaladisailoed T, Numnonda T.Machine learning models comparison for bitcoin price prediction. *10th InternationalConference on Information Technology and Electrical Engineering*. 2018;p. 506–511. Available from: 10.1109/ICITEED.2018.8534911.

[7]. Jaquart P, Dann D, Weinhardt C. Short-term bitcoin market prediction via machine learning. *The Journal of Finance and Data Science*. 2021;7:45–66. Available from: 10.1016/j.jfds.2021.03.001.

[8]. Rane PV, Dhage SN. Systematicerudition of bitcoin price prediction usingmachine learning techniques. *5thInternational Conference on Advanced Computing & Communication Systems (ICACCS)*. 2019;p. 594–598. Availablefrom: 10.1109/ICACCS.2019.8728424.

[9]. Roy R, Roy S, Hossain MN, Alam MZ, Nazmul N. Study on nonlinear partial differential equation by implementing MSEmethod. *Global Scientific Journals*. 2020;8(1):1651–1665.

[10]. Mckinney W. Pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*. 2011;14(9):1–9.Available from:https://www.dlr.de/sc/portaldata/15/resour ces/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf.

[11]. Waskom M. Seaborn: statistical data visualization. *Journal of Open Source Software*. 2021;6(60):3021. Availablefrom: https://dx.doi.org/10.21105/joss.
03021.

[12]. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*. 2014;8:14. Available from:https://dx.doi.org/10.3389/fninf.2014 .00014.

[13]. Polo G. PyGTK, PyQT, Tkinter andwxPython comparison. . *The Python Papers*. 2008;3:26–37. Available from:https://www.mclibre.org/descargar/docs/ revistas/the-python-papers/the-python-papers-3-1-en-200804.pdf.