

# Forecasting Bitcoin Price Considering Macro Economic Factors and Media Influence Using Bidirectional LSTM and Random Forest Regressor as Ensemble Model

by

Swattic Ghose  
19101216

Faiyaz Bin Khaled  
19101138

Nafiz Imtiaz Rafin  
19101169

Rubaiyet Hossain Jawwad  
19101079

Yamin Bin Yahiya  
19101274

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science & Engineering.

Department of Computer Science and Engineering  
Brac University  
January 2023

© 2023. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

## Student's Full Name & Signature:



---

Swattic Ghose  
19101216



---

Faiyaz Bin Khaled  
19101138




---

Nafiz Imtiaz Rafin  
19101169



---

Yamin Bin Yahiya  
19101274



---

Rubaiyet Hossain Jawwad  
19101079

# Approval

The thesis titled “Forecasting Bitcoin Price Considering Macro Economic Factors and Media Influence Using Bidirectional LSTM and Random Forest Regressor as Ensemble Model” submitted by

1. Swattic Ghose (19101216)
2. Faiyaz Bin Khaled (19101138)
3. Nafiz Imtiaz Rafin (19101169)
4. Yamin Bin Yahiya (19101274)
5. Rubaiyet Hossain Jawwad (19101079)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science & Engineering on 19th January, 2023.

**Examining Committee:**

Supervisor:  
(Member)



---

Dr. Md. Golam Rabiul Alam

Professor  
Department of Computer Science and Engineering  
Brac University

Program Coordinator:  
(Member)



---

Dr. Md. Golam Rabiul Alam

Professor  
Department of Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)

---

Sadia Hamid Kazi

Chairperson and Associate Professor  
Department of Computer Science and Engineering  
Brac University

# Abstract

The decentralized cryptocurrency has created many opportunities for secure and safe financial transactions with a bright prospect. The cryptocurrency market rapidly expands, leading to erratic price movements due to geopolitical, social, and other macroeconomic factors. As a result, the price of such cryptocurrencies changes every day. For our research, we limit our scope to predicting and forecasting bitcoin prices accurately. For predicting the trend of Bitcoin price, we considered two major factors: the consideration of various macroeconomic markets and the sentiment analysis of social media. Our contribution to this research was the volume of data that we collected for sentiment analysis for tweets which is approximately 85 millions. In addition, we considered the impact of the markets of AMD and NVIDIA which are the main tech companies that provide consumer level GPU that has a huge impact in cryptocurrency mining, which has never been considered before for predicting cryptocurrency prices and to improve our accuracy we used ensemble Random Forest Regression with Bidirectional LSTM. In this case, we considered Twitter. We have used the Vader Sentiment Analysis model to calculate the sentiment scores (positive, negative, neutral, and compound). We have used four parallel Bayesian Optimized Bi-LSTM models, each with its input features, to combine their predictions and train an ensemble Random Forest Regressor with those predictions. Then, we used the trained RFR model to pick the best forecast out of those four parallel Bi-LSTM models. Furthermore, we got the following results:  $MSE = 0.0021607$ ,  $MAE = 0.0318709$ ,  $R2 = 0.99909$ , and  $MAPE = 0.0038217$ . The findings were that Bidirectional LSTM functions better in prediction when we consider sentiment analysis and other macroeconomic factors (AMD, NVIDIA, S&P 500, NASDAQ, GOLD stock prices). Moreover, using RFR as an ensemble model, the accuracy is boosted significantly.

**Keywords:** Bitcoin, Macro Economic, Sentiment Analysis, GPU, Bayesian Optimization, Bidirectional LSTM, Ensemble Random Forest Regressor.

## **Acknowledgement**

At first, all the praise would go to Almighty Allah for whom our thesis have been completed without any major trouble. The Almighty have blessed us with good health.

Secondly, we are really grateful to our supervisor, Dr. Md. Golam Rabiul Alam Sir for his kind support and with his valuable advice in our research work. He helped us whenever we needed any support.

And finally to our beloved parents, who supported us throughout our life. With their kind support and prayer, we are now on the verge of our graduation.

# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Approval</b>	<b>ii</b>
<b>Abstract</b>	<b>iv</b>
<b>Acknowledgment</b>	<b>v</b>
<b>Table of Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>Nomenclature</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Research Problem . . . . .	1
1.3 Research Objective . . . . .	2
<b>2 Literature Review and Related Work</b>	<b>3</b>
2.1 Existing Work . . . . .	3
2.2 Existing Work Summary . . . . .	12
<b>3 Methodology</b>	<b>14</b>
3.1 Dataset . . . . .	15
3.1.1 Historical Data . . . . .	15
3.1.2 Twitter Data . . . . .	15
3.2 Dataset Pre Processing . . . . .	15
3.2.1 Handling Missing Values . . . . .	15
3.2.2 Feature Extraction . . . . .	16
3.3 Feature Selection . . . . .	20
3.3.1 Correlation . . . . .	20
3.3.2 Data Distribution . . . . .	22
3.4 Model Specification . . . . .	23
3.4.1 Vader (Valence Aware Dictionary sEntiment Reasoning) . . . . .	24
3.4.2 LSTM (Long Short Term Memory) . . . . .	24
3.4.3 GRU (Gated Recurrent Unit) . . . . .	25
3.4.4 Bidirectional GRU . . . . .	26

3.4.5	FB Prophet . . . . .	27
3.4.6	Bidirectional LSTM . . . . .	27
3.4.7	Random Forest . . . . .	28
3.4.8	Overall Model Specification . . . . .	29
<b>4</b>	<b>Results and Discussion</b>	<b>31</b>
4.1	Error Matrices . . . . .	31
4.2	Result Analysis . . . . .	32
<b>5</b>	<b>Conclusion and Future Work</b>	<b>36</b>
<b>6</b>	<b>References</b>	<b>37</b>



# List of Figures

3.1	Top Level Overview of the Proposed System . . . . .	14
3.2	Heatmap of First Three Individual Dataset . . . . .	16
3.3	Heatmap of Last Three Individual Dataset . . . . .	17
3.4	Bar Graph of Tweet Counts Annually . . . . .	18
3.5	Visual Comparison of Tweet Counts Annually . . . . .	18
3.6	Exact Tweet Counts Annually . . . . .	19
3.7	Correlation Matrix . . . . .	20
3.8	Covariance Matrix . . . . .	21
3.9	Scattered Plot of Features with Respect to Label . . . . .	22
3.10	Kernel Density Plot of Open . . . . .	23
3.11	Box Plot of Open . . . . .	23
3.12	LSTM Cell . . . . .	25
3.13	LSTM Cell Details . . . . .	25
3.14	Random Forest . . . . .	29
3.15	Overall Model Specification . . . . .	30
4.1	Performance score of the four models . . . . .	33
4.2	Prediction of all four models . . . . .	34
4.3	Final Forecast Graph for Ensemble Model . . . . .	35

# List of Tables

4.1	Comparative Analysis of Multiple Time Series Models . . . . .	32
4.2	Performance score of the four models . . . . .	33
4.3	Prediction of all 4 models . . . . .	34

# Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

*BiGRU* Bi directional Gated Recurrent Unit

*BiLSTM* Bi directional Long Short Term Memory

*BTC* Bitcoin

*FBProphet* Facebook Prophet

*GRU* Gated Recurrent Unit

*LSTM* Long Short Term Memory

*MAE* Mean Absolute Error

*MAPE* Mean Absolute Percentage Error

*MSE* Mean Squared Error

*R2* R-Squared

*RFR* Random Forest Regressor

*Vader* Valence Aware Dictionary sEntiment Reasoning

# Chapter 1

## Introduction

### 1.1 Background

Cryptocurrency was created with the intention of decentralizing money and facilitating transfers without the aid of banks. American cryptographer David Chaum created the first encrypted electronic currency known as Digicash in 1995. In 1998, Nick Szabo originally developed Bitgold, the ancestor of the present Bitcoin [37]. Bitcoin's price changes a lot, compared to other monetary assets (such as gold, stocks). For instance, Bitcoin's price can occasionally fluctuate by more than 10% (or even 30%), and its volatility is far higher than that of gold[39], crude oil [38], and S&P 500 equities [41]. Due to these unusual and significant price movements, investors and academics are curious to understand more about them in order to predict the direction of the market [42]. It is important to consider all the factors that would impact the bitcoin price. We have considered the Open, Low, High and Volume features from the bitcoin prices that will be fed into our models [43]. We also deeply studied the media impact on bitcoin prices [43], so sentiment analysis was done by fetching relative data related to bitcoin from Twitter . We also found that GPU is required in bitcoin mining, so we also analyzed the GPU market (NVIDIA and AMD stock prices) and its relation with the bitcoin prices [45]. Lastly, we considered economic factors which can relate to bitcoin price changes [40]. Our research would focus on the external factors (economical factors & social media influence) that would influence the bitcoin price and subsequently develop an ensemble model using Bi directional LSTM (Long Short Term Memory) model and Random Forest Regressor which will be used for forecasting the bitcoin's "Open" price with higher accuracy.

### 1.2 Research Problem

The high price volatility of cryptocurrencies is a major drawback as it demotivates investors from making investments in the bitcoin market, despite the fact that they have a high potential to function as financial assets for businesses and as electronic currencies for more secure and decentralized transactions. Moreover, there can also be other underlying factors that would affect the price of Bitcoin. Bitcoin's price trend also fluctuates along with the other market values, such as stock values. It is not always possible to understand the trend of bitcoin price based on its previous values. Furthermore, it is also important to develop predictive models with better

accuracy. It is important for investors to find reliability on the model that would be used to predict the price trend of bitcoin.

### 1.3 Research Objective

Our research objective is to deeply study the volatility nature of bitcoin's price. Our goal is to study the economic factors that would influence bitcoin's price. We would also analyze how social media would impact bitcoin's price and study the correlation between them. For the media influence part, we would analyze the sentiments of tweets regarding bitcoin and study its impact on bitcoin price. We have collected a large volume of tweets, approximately 85.5 million of tweet data was collected for our research purpose. We would be assessing the market stock prices and deeply understand the impact on bitcoin's price trend. Along with the stock prices, we have also analyzed the GPU stock prices and its impact on bitcoin's price. Our aim is to build a forecasting model with more reliability and including more global market factors. We would be developing a multivariate forecasting model in order to consider all the external factors affecting the bitcoin's price trend. The main objective of this paper are:

1. We collect our historical datasets of Bitcoin and other stock prices of various companies and resources from Yahoo! Finance and perform thorough data cleaning to make it readable by our model.
2. We collect a huge quantity of Twitter tweets by scraping method and perform Vader Analysis to produce compound scores.
3. We combine our data in one file and perform different correlation and co-variation to select the features which are relevant in predicting our selected label.
4. We have also considered the impact of GPU stock prices (NVIDIA and AMD) on bitcoin price.
5. We also considered the impact of macroeconomic factors on bitcoin such as gold and stock prices.
6. We create 4 Bi-LSTM models with separate factors and produce predictions which we later feed into the Random Forest Regressor to create our ensemble model.
7. We compare our 4 models and the ensemble model with multiple error metrics to show statistical validation.
8. We generate a comparative analysis with other time series models and show that our ensemble model outperforms all the time series models in comparison with a high margin.

# Chapter 2

## Literature Review and Related Work

### 2.1 Existing Work

According to the research paper [1], the unpredictability of various trends and factors that affect the price of the cryptocurrency (Bitcoin, Ethereum, Ripple) value information can be analyzed and made into a more predictable model by the means of Deep Learning algorithms. Long Short-Term Memory(LSTM)is used and to analyze the price dynamics of the cryptocurrency, Artificial Neural Network(ANN) is used. In their study, the prediction made by the ANN model with a memory of previous 1 month data outputs a better result in forms of Mean square error and Pearson correlation whereas LSTM performs well in predicting the price of cryptocurrency 1 day into the future when the train data used is of a small time frame(1 day or 3 days). The research does have a strong base in understanding the trend in prices by observing the past data of cryptocurrencies but does not take into account different hidden factors. It also generalizes the cryptocurrency market with only 3 currencies which does not provide the viable grounds for proper market analysis. Furthermore, the research does not provide any means of data optimization which might increase the efficiency of the output predicted from the data.

According to the research paper [2], the price volatility and unstable behavior, the cryptomarket investment does not produce the desired profits. The authors developed a statistical approach based on the Random Walk theory, which has proven to be quite useful in financial research, to forecast the real-time price of bitcoin. For currencies like Bitcoin, Ethereum, and Litecoin, their methodology uses Multilayer Perceptron (MLP) and LSTM. Different literature has been challenged in the study for favoring a deterministic approach to price prediction over a stochastic approach. The authors made a comparison and contrast between several current research approaches to the problem and their proposed model. The writers supplied their results to back up their assertion, and this definitely speaks in their favor. According to the journal article, the average relative improvement of stochastic based neural networks over regular neural networks in the Norm dataset is 1.56 percent when the perturbation factor is at range 0:1. There are also unknown factors that influence bitcoin prices, such as market inefficiencies caused by information asymmetry. Random walk theory may not be a useful statistical approach for predicting bitcoin price in this scenario. To solve this problem, the hidden elements that influ-

ence the pricing must be taken into account.

The authors of the research [3] used noise-correlated stochastic differential equations to develop a framework that discussed cryptocurrency price fluctuations and created a correlation with social media activities. Geometric Brownian is employed to train bitcoin rates in this study, while the Geometric Ornstein process is used to model social media activities and cryptocurrency trade volume. Their methodology allows for a better grasp of the likelihood of success when investing in bitcoin in a short period of time. The findings of this research are quite true, with a higher MAPE score. The authors claim that their approach can forecast data in a three-month time frame (April, June, August). Their conclusion was based on the notion that the price dynamics of cryptocurrencies and the stock market are identical. The authors used data from the DARPA SocialSim project to compile their findings. The dataset covers data from the 17th of January to the 31st of August 2017 and is based on the premise of time independent parameters, implying that the model's correlations and parameters are not time variables. The authors found from their research that the extracted empirical noise of diverse time series is associated, leading to the development of a framework that can make predictions to some extent, is robust, and is easily generalizable. The accuracy of the model, according to the authors, is determined by the underlying equation of the various time series. The authors also discuss the study's shortcomings and limits. When the model tries to predict the price of a crypto currency more than one day in the future, it encounters multiple uncertainties, according to this study. Despite this, their model was able to match the original price data's primary variables and attributes. The authors further point out that because this model is very generic, it makes few assumptions about the type of information. The authors' most significant drawback is the independent prediction technique, which requires the model to generate a correlation to known data in order to quantify predictions. In this paper [4], the authors approached the prediction of Bitcoin price fluctuations through orthodox trading indicators, features produced through De-noising and basic features which are evaluated by Attentive LSTM Embedded Network (ALEN). Both of these techniques have their own perks. For example, the time dependency representation of Bitcoin price can be acquired through LSTM. On the other hand, hidden representations from similar cryptocurrencies can be acquired through an embedding network. It has been proven through experiments that out of all the baselines better performance can be achieved through an Embedding Network. Unfortunately as related data regarding Bitcoin is still limited due to active trading since only in 2017, it won't be possible to integrate large volumes of data. For this reason minute level data were integrated in this research where mining information has been ignored. This research study [5] presents a model for forecasting the price of the well-known cryptocurrency Bitcoin using various neural network methodologies, including Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM). This study compares the proposed model against various existing models, such as RNN with LSTM, Linear Regression, and Random Forest, that have been applied to the same domain. This suggested technique yields an MAE (Mean Absolute Error) of 0.0043s, which is considerably less than what the Random Forest and Linear Regression models provide. However, this article clearly disregards the fact that such a turbulent market cannot be accurately anticipated by considering only one cryptocurrency. There are other additional prominent cryptocurrencies whose market trends should be evaluated. In addition,

hidden variables such as social media influence likely to have a considerable impact on the cryptocurrency market.

In this journal [6], the authors proposed a deep neural model based on a multi input architecture based on a study that was performed using the data of three popular cryptocurrencies which are most often in the lead in terms of market capitalization. Three consecutive years of data from the website CoinMarketCap were taken into account to conduct the research using the model MICDL(Multiple-Input Cryptocurrency Deep Learning Model). The main motive behind using this specific model was to extract useful crypto data and subsequently process them for achieving highly accurate predictions. Overall this paper is really solid in terms of experimental analysis, as it is seen that the model was able to properly develop a forecasting model which not only has the capability to mix different cryptocurrency data effectively and secure lower computational cost compared to a traditional model, but also has the best classification and regression performance. However, more sophisticated pre-processing techniques like exponential smoothing and moving average can be applied on the MICDL model to yield better results. Besides this, recent market analysis dictates that social media plays an integral role in the cryptomarket which should also be considered.

In this research paper [7], the authors have employed both linear and non-linear error correction models to forecast the market trend of any cryptocurrency. To better and accurately predict its price, linear ECM along with granger casualties were used for 14 different cryptocurrencies to better understand how other cryptocurrencies affect BTC. Linear ECM is used as its a far better model compared to other NN and autoregressive models when it comes to RMSE, MAE and MAPE. After careful data analysis, it was found that, with highly correlated cryptocurrencies, linear ECM can be used to predict future log-return values of each coin. However, the cryptomarket is extremely volatile and has a wide range of factors for fluctuations. Therefore to accurately predict the market, hidden factors such as spread of misinformation, influence of social media etc should be considered, as they play an integral role in the fluctuation of the crypto market.

The authors in this journal[8] for anticipating the market price of the cryptocurrencies including Bitcoin, Litecoin and Ethereum have used three Recurrent-Neural-Network models which are Gated Recurrent Unit(GRU), Long short-term memory(LSTM) and bi-directional LSTM. To compare results, Mean absolute percentage error(MAPE) was used for prediction. Among the three RNN models, GRU showed the best results and the results are 0.2454% for Bitcoin 0.8267% for Ethereum and 0.2116% for Litecoin. But factors such as trading volume, social media and tweets were not taken into account which could potentially affect the prices of the crypto currency.

The authors in the journal [9] have demonstrated the performance of Random Forest model on data of Bitcoin and Ethereum historical data at a 5 minutes interval from Binance and Binfindex datasets. The author made two analyses one is the Accuracy Variation and another is Factor importance. In the accuracy variation we see that the accuracy initially increases prior to becoming stable which is up to 70% accuracy as the time interval increases thus the prediction accuracy of the model increases with future ahead. For the Factor importance analysis the author chose 16 factors in Alpha101, among which alpha24(above 0.25) and alpha32(above 0.2) which indicates significant level of predictability in the analysis, also features such as



close, high, low has degree of predictability of above 0.15 , 0.1 and 0.1 respectively. The research gap that was reflected was using only market data to build the data and not considering the factors : the position and status of the cryptocurrency in the market, cryptocurrency community current actions, social media trend.

The authors in the paper [10] have proposed HMM (Hidden-Markov-Models) to describe cryptocurrencies previous market trends along with the LSTM model to calculate future trends which turned out to work better than ARIMA, which is a traditional time-series prediction model) as well as the LSTM. For data the author chose 2-minute interval Bitcoin data from Coinbase transaction market. The performance metrics that have been used are MSE, RMSE and MAE. The results for HMM-LSTM are 33.888(MSE), 5.821(RMSE), 2.510(MAU), for ARIMA are 20153.722(MSE) 141.964(RMSE) 112.060(MAU), for traditional LSTM are LSTM 49.089(MSE) 7.006(RMSE) 2.652(MAU). The author states more additional features could have been extracted from the bitcoin transaction which could provide more insightful information.

According to Journal [11], forecasting the bitcoin market is a challenging endeavor because of the highly stochastic nature of the market trend. The study explains how a new feature, interactivity, dramatically enhanced the price prediction model based on the historical price. Gradient Boosting Algorithm is superior to machine learning models since optimization is based on function search space rather than parameter search space, according to the article. VADER(Valence Aware Dictionary for Sentiment Reasoning), introduced by Hutto and Gilbert, is a well-known 'rule-based' paradigm primarily required in sentiment analysis, according to the publication. The emotion spans from -4 to +4, where -4 represents severe negativity and +4 represents extreme positivity. The VADER gets a document as input and calculates its sentiment score. Sometimes, heuristic criteria are introduced to demonstrate the sentiment value. The key weakness of this work was its reliance on Twitter and Bitcointalk for data collection. Bitcointalk's data availability was quite limited. The investigation was done just on bitcoin; other cryptocurrencies were not examined. Finding a reliable way to calculate the daily sentiment score was the final barrier. In the article [12], the authors Darapaneni N. , Anwesh R. Paduri, Sharma H. , Hindelkar N. , Aiyer U. and Agarwal Y. predict the stock price by combining the stock prices with sentiment scores using LSTM, Bi-LSTM, Linear Regression, Arima, KNN, FbProphet and Random Forest. Other parameters such as Gold prices, Oil prices, USD exchange rate and Indian Govt. Securities yields were considered as a feature for their models. The stock prices were collected from Yahoo Finance, and features such as Open, High, Low and Close were selected for the model training. They could not collect data from Twitter due to accessibility, so they only considered online News sources such as Hindustan Times, News18 and Mint. They used a Sentiment Intensity Analyser to calculate the sentiment scores. In the first part, they only gave input features related to the stock price dataset to check which model worked better in their favor. They compared the RMSE score of different models and they found out LSTM gave better accuracy compared to the other models. In the second part, they have considered other features such as Gold prices, Oil prices, Exchange Rates along with sentiment scores that were given as input into the Random Forest model for prediction. Later, they plotted a prediction graph of the 2 different models. One of the challenges of this paper was collecting relevant news from online. Another challenge was collecting data in the same date range.

They could not implement multivariate LSTM which might have given better results. In the article [13], the authors Inamdar A. , Bhatt S. , Pooja Shetty M. and Bhagtani A. have predicted the price of cryptocurrency by sentiment analysis. For their research purpose, they collected tweets and news feed for calculating the sentiment scores. The twitter data is collected using web scraper and news data is collected from webhose.io. They have omitted those data which were not in the English language. For sentiment analysis, they have used RNN along with LSTM. RNN helps to remember short patterns and LSTM has memory cells which store long sequences of data. They have built multi-class classification using keras library to calculate the positive and negative score. Then the sentiment scores are merged with bitcoin prices. For prediction, they have used a Random Forest regressor. They have included a bootstrapping technique which in turn generates multiple decision trees. This would overcome the problem of overfitting. Their result for the first day prediction accuracy score is MAE=2.7526 and RMSE=13.70. The second day prediction accuracy score is MAE = 3.1885 and RMSE=15.1686.

In the article[14], the authors Serafini G. , Ping Y. , Zhang Q. , Brambilla M., Wang J. , Yiwei H. and Beibei L., have predicted the price of bitcoin by analyzing the market along with the sentiments related to bitcoin. In their research work, they have analyzed the financial and sentiment features gathered from economic and online sources. They have applied two models which are ARIMA and RNN and later established a comparative study between the models. In their research, they have calculated the average price of bitcoin per day and they discarded the features such as High,Low, Open and Close. They only considered weighted average and volume as input features along with the sentiment scores. They have collected tweets using web scraper and then calculated the number of tweets related to bitcoin per day. They used Vader to calculate the sentiment scores. For their ARIMA model, they first applied the Dickey Fuller Test in order to determine whether the data is stationary or not. They estimated the parameters of ARIMA using Maximum Likelihood Estimation (MLE). Their second model was adding a LSTM with RNN as LSTM provides a memory cell to it. Later, they calculated the overall performance of the two models where RNN has outperformed the ARIMA model.

In the article [15], the authors Cakra Y. and Trisedya B. have predicted the price of stocks using Linear Regression models and developing relationships with the sentiments related to the stock market. Their main objective was to analyze the Indonesian stock market and its other factors which may influence the price of stock. For their research purpose, they have collected the data from Twitter which are related to the stock market. Naive Bayes and Random Forest are used to classify the sentiment scores. The stock dataset was collected from Yahoo Finance. The tweets were classified into 3 classes (positive, negative and neutral). They discarded the neutral tweets because they considered those as spam tweets. A linear regression model was established to feed the stock prices along with sentiment scores in order to predict the prices of future days. From their research work, Random Forest has given better accuracy than Decision Tree, Support Vector Machine and Naive Bayes. They can improve their research by POS tagging and word weighing. They would like to conduct their research on large scale data. Their limitation was not using models other than linear regression models.

In the article [16], the authors Tingwei Gao, Yueting Chai and Yi Liu displayed Stock closing price prediction using NN model which is LSTM in comparison with other

models to evaluate the performances and proved that LSTM works marginally better than the other models which were MA(Moving Average), EMA(Exponential Moving Average), SVM(Support Vector Machine). The stock market that was analyzed was of S&P 500 from data ranging from Jan 3,2000 to Nov 10, 2016 collected from Yahoo Finance. Gao used three layered models out of which the input layer had 6 neural units, LSTM layer 10 and output layer 1, with mesh connection with all the proceeding neural units. For forward and backward propagation activation functions that were used were sigmoid, tanh and ReLU(Rectified Linear Unit). Lastly for optimizing the learning rate and bias, ADAM optimizer and Gradient Descent were implemented respectively. It turned out that LSTM outperformed the other models in terms of MAE, RMSE, MAPE and MPE scores.

In the article [17], Apache Spark's Markov Chain Monte Carlo was used to forecast future stock values. The study shows how different neural networking and complex mathematical computation had been used for prediction but the apache spark makes the computation faster by dividing its work in parallel computing. Apache Spark was created primarily to process massive amounts of data using distributed cluster technologies. Utilizing Apache Spark to aggregate numerous computers speeds up computation.. In order to estimate the likelihood of future stock prices running on the Apache Spark multi node cluster, they analyzed stock prices using technical analysis and Markov Chain Monte Carlo. MCMC, is a technique for sampling from posterior distributions. By building a model from a probability distribution and using Bayesian inference, MCMC approaches simplify the situation. Probabilities are understood as subjective degrees of belief in Bayesian inference . The Bayes' Theorem serves as the basis for the Bayesian inference model. The Markov Chain Monte Carlo (MCMC) technique creates a Markov chain in a stationary special distribution to sample from a distribution. It generally uses two ways to sample : The Metropolis-Hastings Algorithm and Gibbs Sampling. Geometric. Brownian Motion is used by the writers to estimate the cumulative return of stock values across a portfolio. In order to build the model, the authors used PyMC3 framework from python and their test hypothesis was done using the Student T-distribution. The authors processed raw data that came from the Yahoo Finance API. This data was then converted to log files. They had 540 days from the data collected from 2014-2015. They developed the model with PyMC3 to determine the prior, likelihood, and posterior distribution, supposing that the data are normally distributed, but the data were represented as Student-T distribution. They then created a model for data distribution before deciding on the parameter values and validated the model with a posterior predictive check .Following this, they can determine how well their model matched data and determined the likelihood that stocks will rise or fall the following day. They ran tests using combinations of large and small data with large and small sampling and checked the running time. The running time was different in each case and increased with more difficult sampling and data size .From their results, they concluded that models for each company have to be tuned and will be different while the outcome of the forecast is influenced by the chosen data. In addition, the processor capability and number of nodes in the Apache cluster also affects the prediction results while more Markov chain generation requires more RAM. Overall, the research agrees that there are not many machine-learning algorithms which are compatible with the Apache Spark library and hence it will need modification in order to be used with further research with modeling.

In the article [18], the authors Rui Fu, Zou Zhang and Li Li have highlighted how the non-linear and stochastic nature of traffic flow cannot be determined by using traditional models like ARIMA and ARMA. But for traffic control, correct traffic flow prediction is a vital part of the Intelligent Transportation System. In this case, even deep learning methods have failed since the proper selection of deep neural networks couldn't be done. For this reason, to predict the short term traffic flow, the authors have utilized LSTM and GRU. They investigated the accuracy of the model prediction on 50 randomly chosen sensors. In this experiment, performance is assessed after LSTM and GRU have both been trained according to a set of stages. The result of their experiment has shown that LSTM and GRU gave a better outcome compared to ARIMA. It can also be seen that in comparison between GRU and LSTM, out of the total time series, the performance of GRU was superior to LSTM 84

In the article [20], the authors Akshita Gupta and Arun Kumar tried to forecast the load of the power system by implementing three models which are Wavelet-ARIMA, ARIMA and Machine Learning. While the third model uses artificial intelligence and blends historical data with climate patterns, the first two models are time series-based. The Wavelet decomposition is carried out to demonstrate the impact of decomposition on time series forecasting and to evaluate the effectiveness of the various discrete wavelets. Averaging and boosting ensemble methods are used by machine learning to integrate the single regression algorithms. The findings suggest that climatic conditions should be heavily considered while developing load forecasting models because machine learning approaches performed better than time-series algorithms. Additionally, it was found that the averaging kind of ensemble models outperformed the boosting type for the majority of the months. As a result, combining the study's climate trends with the ensemble models yields the highest forecasting accuracy.

In the paper [21], the authors have implemented XGBoost and GRU, also known as eXtreme Gradient Boosting and Gate Recurrent Unit, for their model. The XGB-GRU model uses the extensive feature extraction capabilities of XGBoost to extract the hidden information of multiple control variables in industrial data. The model then extracts timing information from industrial data using the GRU special gating unit. The relevance of XGBoost output variables in guiding real production and fixing the issue of neural network incomprehensibility are discussed. The ability to predict the furnace's temperature demonstrates that the proposed XGB-GRU model is preferable to utilizing a single XGBoost and GRU model.

In the article [22], titled "A CNN-LSTM-based Model to Forecast Stock Prices," X. Liu, C. Wu, and Y. Zhang suggests a hybrid optimization technique for stock price forecasting utilizing a mix of CNN and LSTM. The technique aims to increase the accuracy of stock price forecasts by combining the capacity of CNN to extract spatial information from past stock prices with the capacity of LSTM to capture temporal relationships. Previous research has demonstrated that CNN is an useful tool for optimizing image processing and text classification because it can extract spatial information from data (LeCun et al., 2015). Due to its inability to accommodate temporal dependencies, CNN may not be appropriate for sequential data such as time series (Oord et al., 2016). By combining the strengths of these two methods, the authors hope to overcome their respective limits and produce superior results. The authors conducted trials on the historical stock prices of numerous

companies to evaluate the performance of their hybrid technique. The researchers discovered that the proposed technique produced more accuracy and lower prediction error than regular LSTM and CNN, as well as other hybrid optimization techniques such as LSTM-GRU and CNN-GRU. In addition, they demonstrated that the proposed method could accommodate both short- and long-term stock price dependence. Overall, the paper’s proposed hybrid approach combining CNN and LSTM for stock price forecasting is promising. To evaluate the technique’s generalizability, it would be advantageous for future study to apply it to other financial data, such as currency and commodity prices.

In the article [23], the authors Engr. Muhammad Rizwan, Dr. Sanam Narejo and Dr. Moazzam Javed have made an approach and constructed a model to predict the bitcoin price in US dollars. The Bitcoin price index provides important information. This function can be performed by a long-term memory (LSTM) network and a Bayesian recurrent hierarchical (RNH) neural network. The popular ARIMA method for the prediction of time series is in contrast to the thorough training systems. It was anticipated that deep learning techniques will perform better than the subpar ARIMA forecast. Consequently, they employed the Gated Recurrent Network model (GRU) here to forecast the price of bitcoin.

The paper [30] presents a hybrid method for detecting traffic incidents in real-time using data from vehicle sensors and traffic surveillance cameras. Using a Bayesian optimization approach, the method optimizes the hyperparameters of the random forest classifier and the LSTM network. Bayesian optimization is a global optimization technique that is ideally suited for optimizing costly and noisy black-box functions. It operates by developing a probabilistic model of the objective function based on previous function evaluations, and then employing this model to select the next point to evaluate so as to maximize the expected improvement in the objective function. In the context of the hybrid method presented in the paper, the Bayesian optimization algorithm is utilized to improve the hyperparameters of the random forest classifier and the LSTM network to maximize their performance in recognizing traffic events.

The paper [31] presents a ML approach for predicting the price of Bitcoin through sentiment analysis of Twitter data. The authors classify the sentiment of Twitter postings regarding Bitcoin as positive, negative, or neutral. The anticipated sentiment of these messages is then used as a feature in a second machine-learning model that forecasts the Bitcoin price. The authors assert that their method can outperform conventional technical analysis techniques for predicting Bitcoin’s price. They also report that the RNN is 77.62% accurate at classifying the emotion of Twitter posts. It also reveals a moderate correlation of 0.41 between the rise of adverse sentiments about Bitcoin on Twitter and its subsequent price decline.

The paper [32] presents a mathematical model for predicting the rates of cryptocurrencies and related social media activity. The model is based on stochastic differential equations (SDEs) that describe the evolution of a system over time in the presence of random noise. The authors of the research use correlated SDEs to predict the time-dependent evolution of cryptocurrency prices and social media activity. They then utilize the greatest likelihood method, an optimization technique, to figure out the SDE values which best fit the observed data. The estimated metrics are then utilized to predict the future development of cryptocurrency prices and social media activity.

The paper [33] presents a machine learning-based approach for analyzing the emotions of speech in online social media. To categorize speech emotions as positive, negative, or neutral, the authors combine a probabilistic graphical model known as a partly shared deep belief network (PGCDBN) with a bidirectional LSTM. Recurrent neural networks (RNNs) are useful for processing sequential input, such as speech data. One such RNN is the Bi-LSTM network. A probabilistic model that can depict the relationships between various features in the data is the PGCDBN. The Bi-LSTM-PGCDBN model is trained by the authors using a supervised learning method using a collection of voice data annotated with emotions. In the grid search optimization approach, the model is trained with a variety of different hyperparameter values, and the set of values that produces the greatest performance is chosen.

The paper [34] provides a study that predicts the prices of cryptocurrencies using decision tree and regression approaches. In the study, the authors estimate the prices of cryptocurrencies using decision tree approaches such as the J48 algorithm and Random Forest. On the same dataset, they then employ regression techniques such as Multiple Linear Regression and Ridge and Lasso regression to predict the prices. The authors evaluate their models using a database of historical prices and trade volumes for the ten cryptocurrencies with the highest market capitalization. The paper [35] is a comparison of ML models used to predict the price of Bitcoin. The research examines multiple machine learning models for predicting the Bitcoin price, including deep learning-based regression models (GRU and LSTM), Theil-Sen regression, and Huber regression. MSE and R-Square are two popular metrics for continuous variables. In terms of MSE and R<sup>2</sup>, the results indicate that the deep learning-based regression models GRU and LSTM outperform the other models. With an MSE of 0.00002 and an R<sup>2</sup> of 0.992%, GRU has the best results. However, Huber regression requires significantly less time than LSTM and GRU. The research also emphasizes that the quantity of datasets, parameters, and features employed can alter the accuracy of the models (Open, Close, High, and Low). Moreover, the research warns that these results may not be sufficient to anticipate Bitcoin values due to the fact that a variety of factors, such as social media reactions, national digital currency legislation and laws.

The paper[36] describes a study that analyzes different machine learning methods to forecast the price of Bitcoin. The authors compare 3 various ML models, WaveNet, Recurrent Neural Network (RNN), and other machine learning approaches, such as Random Forest, KNN, SVM, and linear regression. The study describes the findings of a study comparing several machine learning models for predicting the price of Bitcoin using error metrics such as MAE and RMSE. The study reveals that the models tend to perform worse as the prediction gap widens and that models such as ARIMA and SVR perform similarly well in most prediction gaps, except for D30. Additionally, all models tend to perform comparably badly for long-term predictions, which the authors attribute to the rising influence of random elements in the prediction. Additionally, the authors stress the challenge of dealing with time-series data and propose examining the continuous nature of the time series for cross-validation. The authors observed that the WaveNet model performed best among all the models in terms of MAE, RMSE, and R<sup>2</sup>. Specifically, they noticed that the WaveNet model had an MAE of 0.0034, RMSE of 0.0032, and R<sup>2</sup> of 0.982.

## 2.2 Existing Work Summary

The unpredictability of various trends and factors that affect the price of the cryptocurrency (Bitcoin, Ethereum, Ripple) can be analyzed and made into a more predictable model by means of Deep Learning algorithms. The authors developed a statistical approach based on the Random Walk theory to forecast the real-time price of bitcoin. For currencies like Bitcoin, Ethereum, and Litecoin, their methodology uses Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM). The research authors [3] used noise-correlated stochastic differential equations to develop a framework that discussed cryptocurrency price fluctuations and created a correlation with social media activities. The authors claim their approach can forecast data in three months (April, June, and August). Their conclusion was based on the notion that the price dynamics of cryptocurrencies and the stock market are identical. This research paper proposes a model predicting the price of the well-known cryptocurrency Bitcoin by applying different neural network approaches. The authors' most significant drawback is the independent prediction technique, which requires the model to generate a correlation to known data to quantify predictions. In this journal [6], the authors proposed a deep neural model based on a multi-input architecture for predicting the price of any cryptocurrency. Cryptomarket is exceptionally volatile and has a wide range of factors for fluctuations. Hidden factors, such as the spread of misinformation, influence of social media, etc., should be considered, as they play an integral role in the fluctuation of the crypto market. The authors in this journal have used three Recurrent-Neural-Network models: Gated Recurrent Unit (GRU), Long short-term memory (LSTM), and bi-directional LSTM. GRU showed the best results among the three RNN models, which are 0.2454% for Bitcoin, 0.8267% for Ethereum, and 0.2116% for Litecoin. However, factors such as trading volume, social media, and tweets were not considered, which could affect cryptocurrency prices. The authors in the paper [10] have proposed HMM (Hidden-Markov-Models) to describe cryptocurrencies' previous market trends. Based on the historical price, the paper describes how a new interaction feature improved the price prediction model significantly. The paper has suggested that Gradient Boosting Algorithm is far better than machine learning models because the optimization is based on function rather than parameter search space. In the article [12], the authors predict the stock price by combining the stock prices with sentiment scores using LSTM, Bi-LSTM, and Random Forest. The stock prices were collected from Yahoo Finance, and features such as Open, High, Low, and Close were selected for the model training. Other parameters include Gold prices, Oil prices, the USD exchange rate, and Indian Govt. Securities yields were considered as a feature for their models. In the article [13], the authors Inamdar A. Bhatt S., Pooja Shetty M., and Bhagtani A. have predicted the price of a cryptocurrency by sentiment analysis. They have built multi-class classification using the Keras library to calculate the positive and negative scores. For prediction, they used a Random Forest regressor. Their result for the first-day prediction accuracy score is MAE=2.7526 and RMSE=13.70. Random Forest has better accuracy than Decision Tree, Support Vector Machine, and Naive Bayes. They can improve their research by POS tagging and word weighing. The stock market that was analyzed was of SP 500 from data ranging from Jan 3, 2000, to Nov 10, 2016, collected from Yahoo Finance. Apache Spark was created primarily to process massive amounts of data using distributed

cluster technologies. The authors processed raw data from the Yahoo Finance API and turned it into log files. They then used a Markov Chain Monte Carlo (MCMC) technique to build a model for data distribution before deciding on the parameter values and validating the model with a posterior predictive check. XGBoost and GRU bring out various controlled variables in industrial data. The timing information in the industrial data is then extracted by the model using the particular gating unit of GRU. A hybrid optimization technique for stock price forecasting utilizing a mix of CNN and LSTM has been proposed. Previous research has demonstrated that CNN is a valuable tool for optimizing image processing and text classification. LSTM is a recurrent neural network that deals with temporal relationships and captures long-term dependencies in sequential data. By combining the strengths of these two methods, the authors hope to overcome their respective limits and produce superior results. The researchers conducted trials on numerous companies' historical stock prices to evaluate their hybrid technique's performance. The support of massive financial data allows the implementation of machine learning algorithms. This paper proposes using GA for the feature selection. Researchers have proposed a deep-learning model using a bidirectional LSTM network implemented on Keras to predict future stock market indexes, taking into account factors such as news and rumors. The model uses a Seq2Seq approach, which can be applied to various tasks involving text and sequence generation. The use of a Long Short-Term Memory (LSTM) network, a type of deep recurrent neural network, can enhance the memory capacity of the model and outperform baseline models on stock datasets. Finally, population-based genetic algorithms (GA) can effectively solve problems related to noise and collinearity by using a series of stages including solution encoding, fitness evaluation, termination checking, selection, crossover, and mutation. High-performing chromosomes are more likely to be selected, while low-performing ones are eliminated.



# Chapter 3

## Methodology

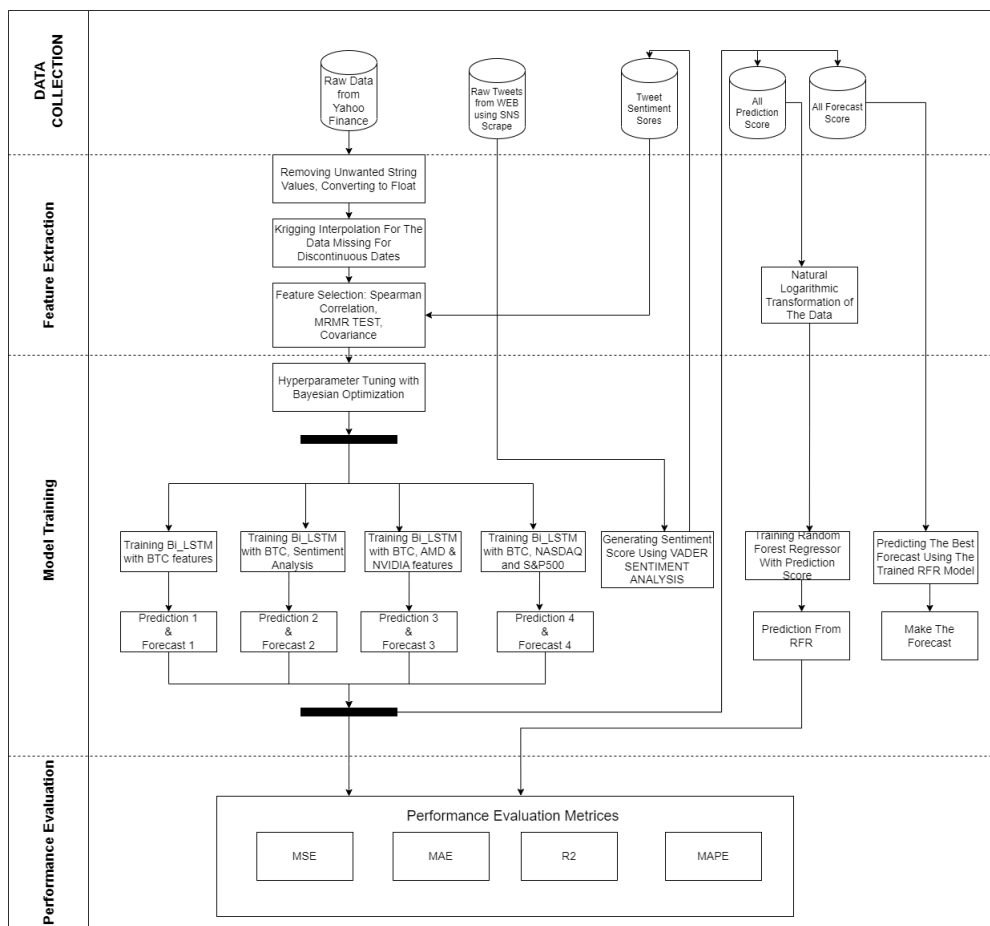


Figure 3.1: Top Level Overview of the Proposed System

We were able to acquire historical stock price information from Yahoo! Finance for Bitcoin, Gold, Crude Oil, AMD, NVIDIA, the S&P 500, and NASDAQ. For the same time period, we used a web scraper called SNS Scraping to collect tweets. After that, the data were cleaned by removing the unwanted characters and converted strings to float values, and then the Kriging interpolation method was used to fill in the values that were absent. The historical data was merged with the Twitter

data. Correlation, Covariance and MRMR was applied to select the features. The hyper parameters were tuned using Bayesian Optimization. Separate the dataset into relevant sets taking into the consideration of the macroeconomic and media influence. Feeding the features into our model and training four parallel Bi-LSTM models to generate prediction and forecast. Saving the prediction and forecast values into separate dataset and later merging them into a single forecast. Using Natural Logarithmic Transformation in the prediction dataset to minimize the impact of outliers. Then, the merged prediction dataset would be fed into the Random Forest Regressor to generate our final prediction. Using the forecast dataset to make a forecast into the future prediction using the trained Random Forest Regressor Model. Finally, we would evaluate the ensemble model using various performance evaluation metrics.

## **3.1 Dataset**

### **3.1.1 Historical Data**

For our research we have collected our Bitcoin, Gold, NVIDIA, AMD, SP500, Crude oil and NASDAQ historical stock data from Yahoo! Finance which gives a more reliable and novel source of information and data. Yahoo Finance is a popular tool when it comes to giving reliable stock prices of companies around the globe.

### **3.1.2 Twitter Data**

For our Social Media Dataset collection, we have used python's library which is 'snsrape'. This library is a web scraper which would scrape social media sites such as Twitter, Facebook. We have set a query where it was instructed to collect any tweets related to bitcoin within our set timeframe. We have considered Twitter for our media influence analysis. The total number of tweets collected are 85563040 which is from 1st January 2015 to December 19, 2022. For our research purpose, we will be using the tweets till 30th August, 2022 which will be fed to our models for predicting the bitcoin opening price.

## **3.2 Dataset Pre Processing**

### **3.2.1 Handling Missing Values**

In our dataset, initially, all the data entries in each row were categorical(Object Datatype). So we converted the values to numerical values. While doing so we found that there were some unwanted characters(char = ',') in between each value which was not enabling us to convert the data. So we removed the unwanted character from each value and then converted the data to numerical values. Also, the empty cells were initially filled with unwanted characters as well. So we removed those as well and proceeded to imputation. For our research we used Kriging interpolation technique. With this approach, a statistical model is fitted to the data in order to estimate the function's value at every given point. Kriging is an effective interpolation technique that works well for issues with vast volumes of data and intricate patterns. In addition, the Bitcoin dataset and Tweets dataset is continuous

with respect to time (all dates have values). But other datasets have discontinuity inside (No market transaction during the holidays). To get rid of this discontinuity, we created those missing dates and set the features as null values and later on imputed them with Kriging Interpolation.

### 3.2.2 Feature Extraction

To find the correlation between each feature of all datasets we used Spearman Rank Correlation. The statistical dependence of 2 variables is measured using this method. It is a non-parametric measurement that is based on the ranks rather than the values of the data.

The Spearman rank correlation coefficient is denoted by the symbol rho ( $\rho$ ) and is defined as:

$$\rho = 1 - (6 \sum d^2) / (n(n^2 - 1)) \quad (3.1)$$

where  $d$  is the difference between the ranking value of the two variables,  $n$  is the number of data points, and  $\sum(\sigma)$  indicates a sum over all the data points.

From our initial data analysis from spearman correlation from each of the Dataset, we have seen some obvious multicollinearity for which we have decided to select one of the features from the list of features that has a very high correlation between each other. The heatmap image generated by the Python Seaborn library gives us insight about multicollinearity.

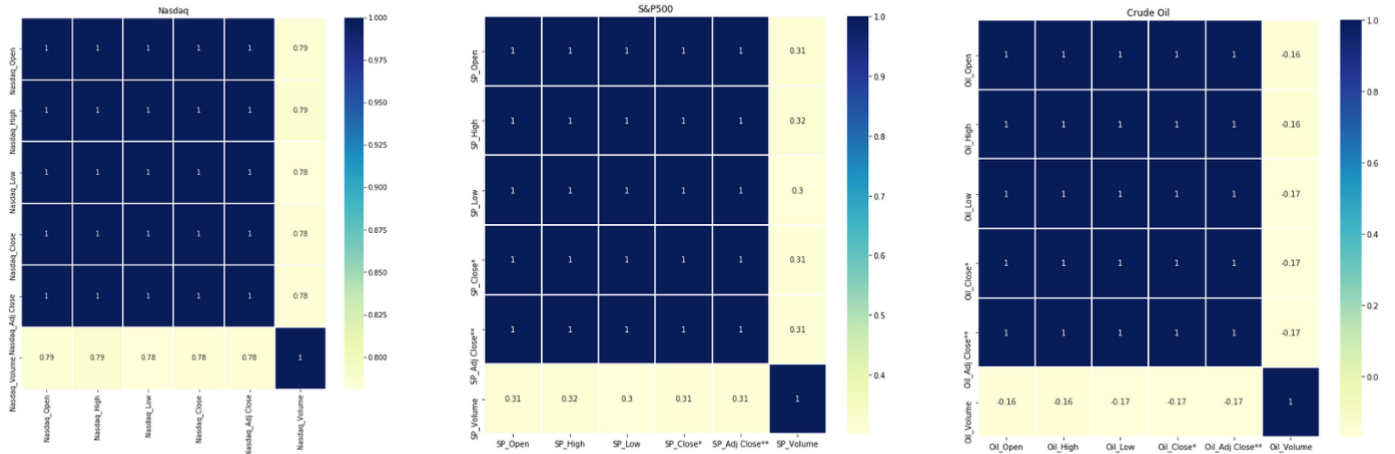


Figure 3.2: Heatmap of First Three Individual Dataset

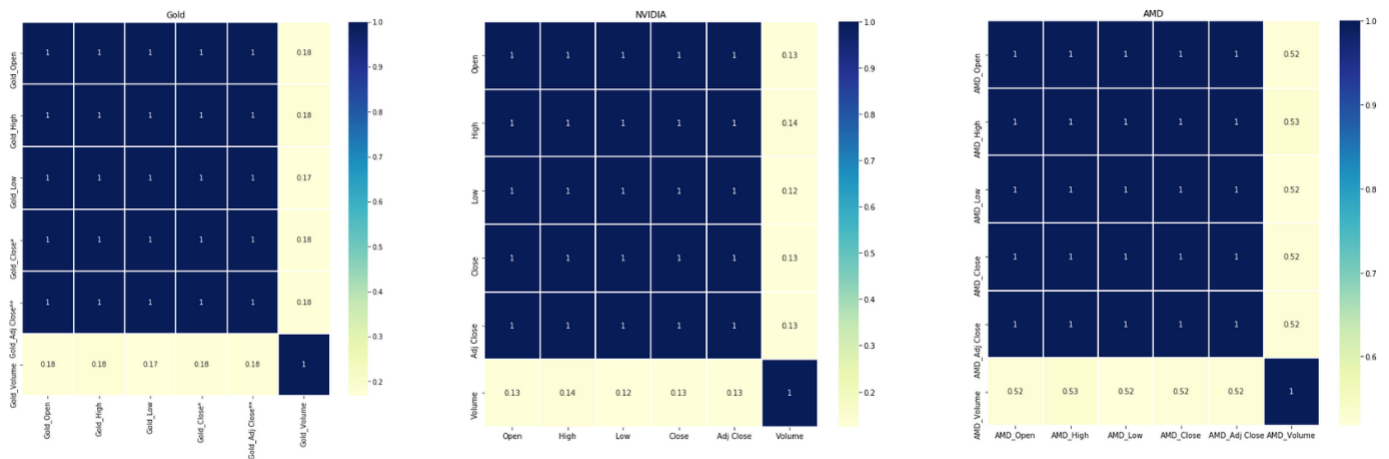


Figure 3.3: Heatmap of Last Three Individual Dataset

From the heatmaps, we can see that all datasets have the common features “Open”, “Close”, “High”, “Low”, “Adj Close” and “Volume”. From the correlation, we can see that all the features except for “Volume” have a correlation of 1 with each other. “Volume” has a correlation below 0.6 which is below the standard threshold of 0.7 for correlation which is why we did not select the feature. Also multicollinearity can be seen among all the other datasets, so to avoid feature redundancy we select only ‘Open’ as the feature to feed into our model.

The total number of tweets collected are 85563040 which is from 1st January 2015 to December 19, 2022. For our research purpose, we will be using the tweets till 30th August, 2022 which will be fed to our models for predicting the bitcoin opening price. After collecting the tweets, we have used Vader Sentiment Analysis to figure out the scores of each text. We have generated 4 scores which are positive score, negative score, neutral score and compound score. Compound score defines the overall polarity of each sentence. Then, we calculated the average, minimum and maximum scores for each day and computed them all into another dataframe along with the number of tweets per day.

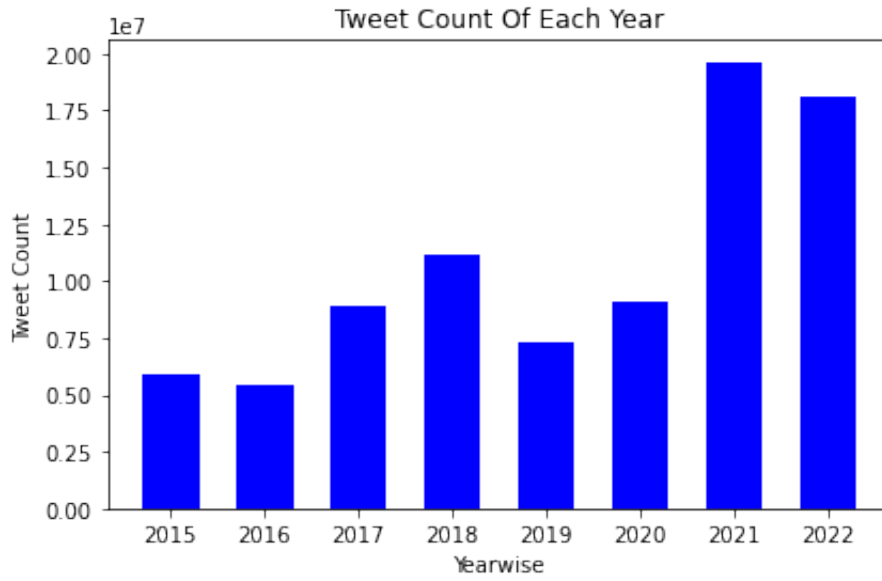


Figure 3.4: Bar Graph of Tweet Counts Annually

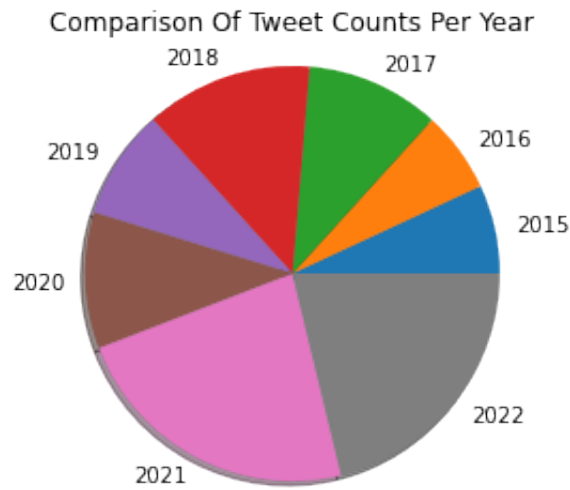


Figure 3.5: Visual Comparison of Tweet Counts Annually

**The No. of Tweets Per Year:**

Year	2015	2016	2017	2018	2019	2020	2021	2022
Tweet Count	5916839	5426381	8918905	11168771	7284706	9067513	19626559	18153366

Figure 3.6: Exact Tweet Counts Annually

### 3.3 Feature Selection

#### 3.3.1 Correlation

For our research, we have selected “Open” of Bitcoin as our target label and now we have to find the relevant features from the dataset that we can use to feed in our model. For that we have used the SelectKBest and mutual\_info\_regression package from the Python sklearn library. 'SelectKBest' is a feature selection method that can be used to figure out a part of a set of the most important features in a dataset. It works by scoring each feature using a function such as mutual information and selecting the top k features based on their scores. 'mutual\_info\_regression' is a function that can be used to compute mutual information between two random variables. In the context of 'SelectKBest', it is used as a scoring function to measure the informativeness of a feature with respect to a target variable. Both 'SelectKBest' and 'mutual\_info\_regression' are part of the 'sklearn' library in Python and are commonly used in machine learning pipelines for feature selection and model building. The images below show the correlation and covariance of our selected label with the rest of the features. For the correlation, we used Spearman Correlation.

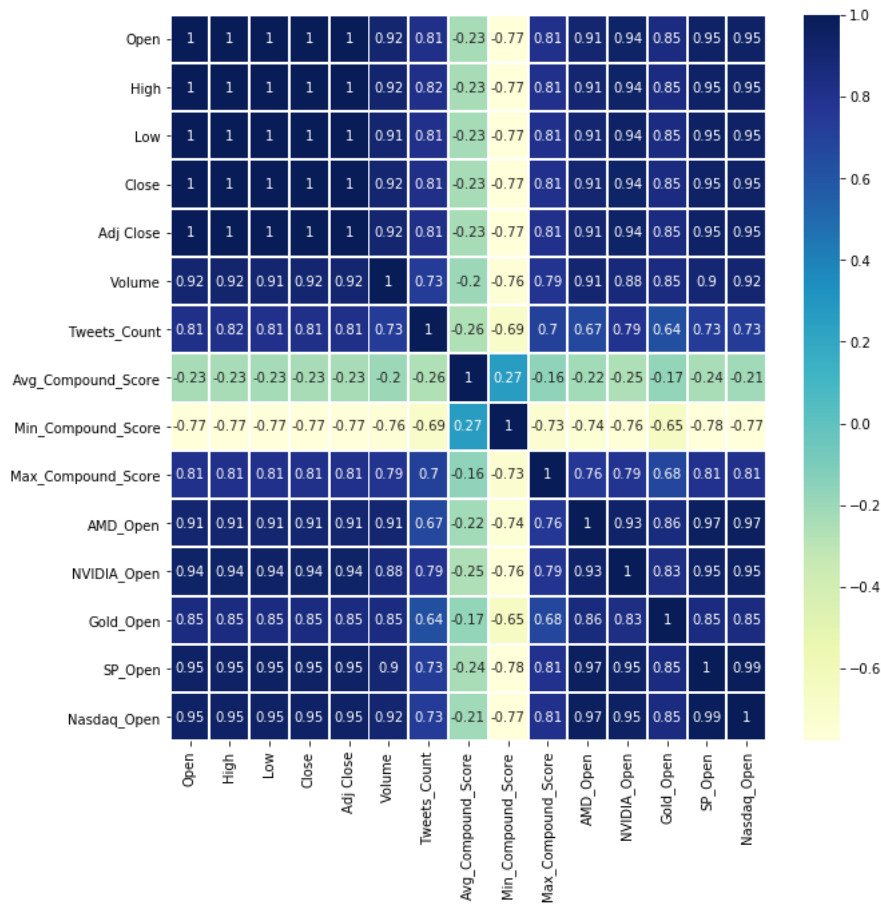


Figure 3.7: Correlation Matrix

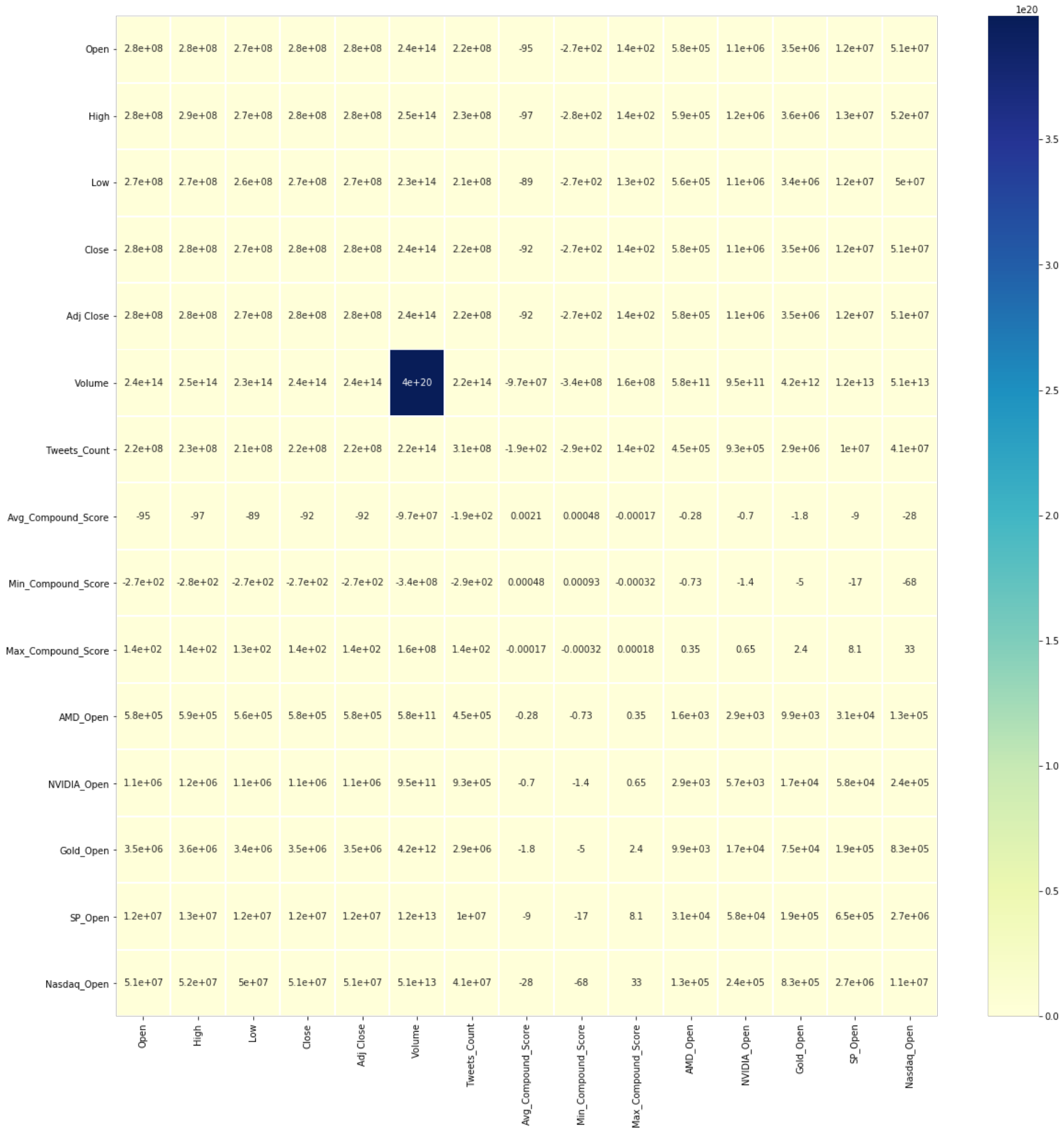


Figure 3.8: Covariance Matrix



### 3.3.2 Data Distribution

Furthermore, if we try to observe the relationship between the selected label and rest of the features, the following Scatter Plot image shows how the data is scattered with respect to the label

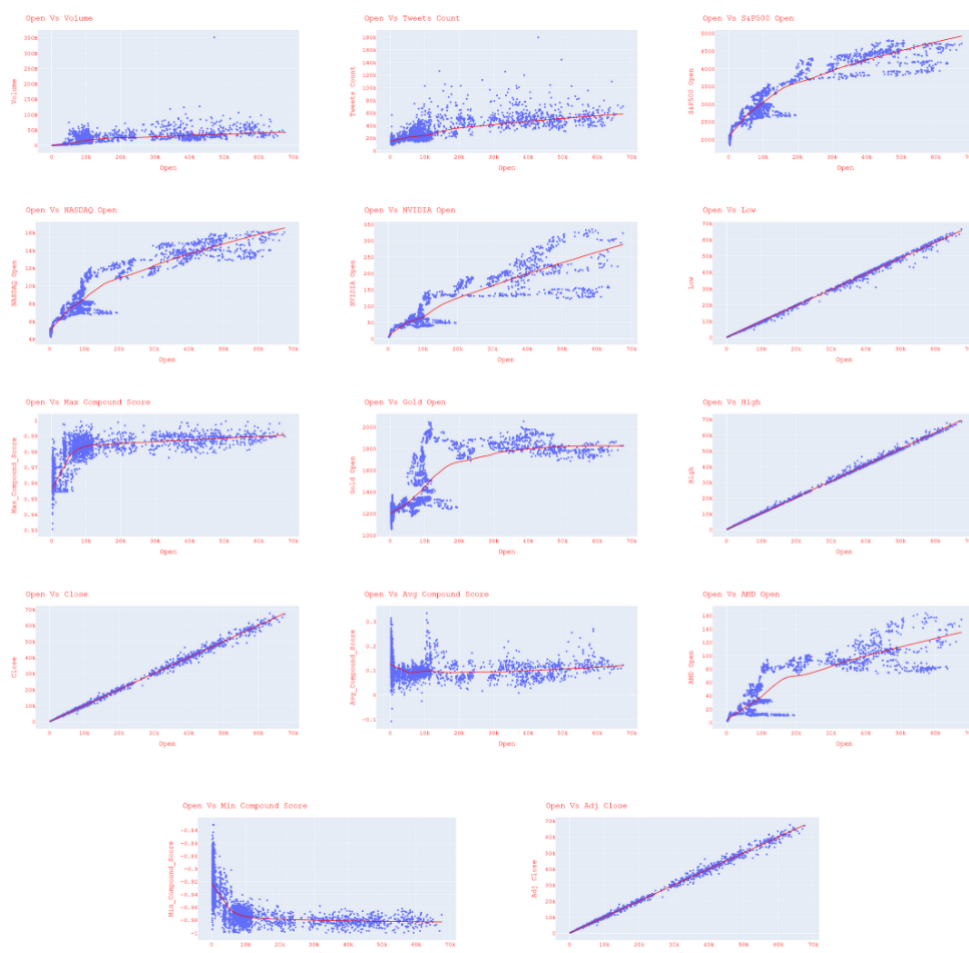


Figure 3.9: Scattered Plot of Features with Respect to Label

From the scatter plot, we can see some wide variations in between the features and label. Features Low, Adj Close, High and Close show clear linear relationship while Min Compound Score and Max compound score shows logarithmic relationship. Other features show a positive monotonic and linear relationship. Furthermore, if we try to observe the relationship between the selected label and the rest of the selected label, the following Scatter Plot image shows how the data is scattered with respect to the label.

Looking at the distribution, we can see that our data is left skewed since the mean is lower than the median. Also the difference between the upper and lower quartile is very high indicating that the data is widely dispersed which is shown in the table.

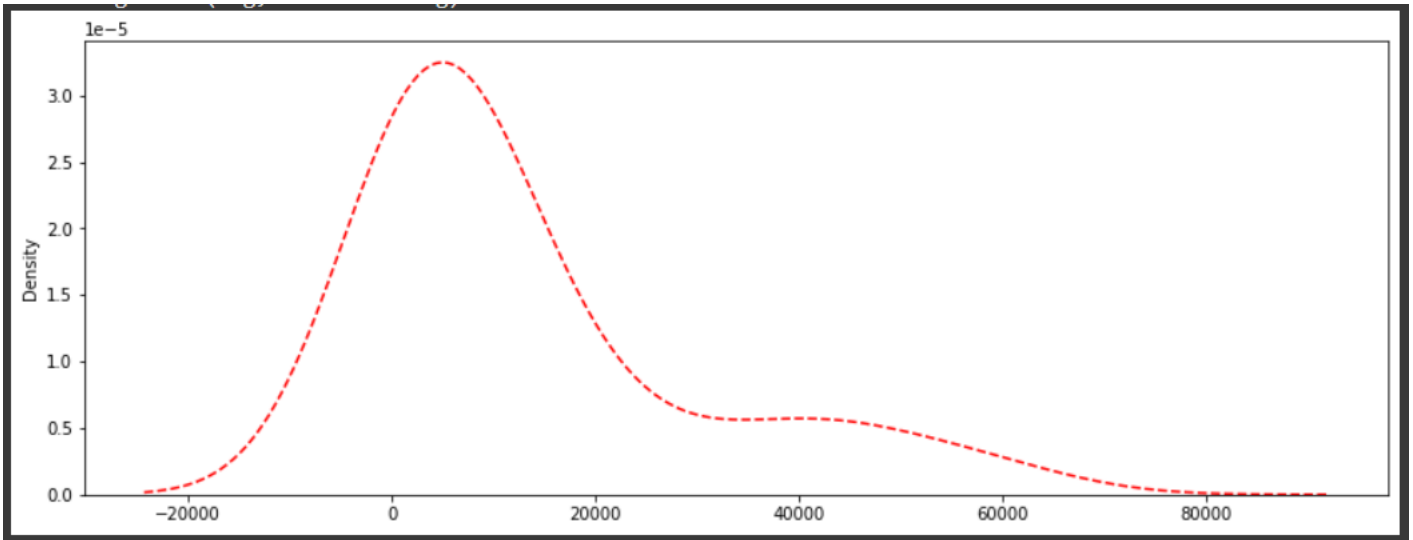


Figure 3.10: Kernel Density Plot of Open

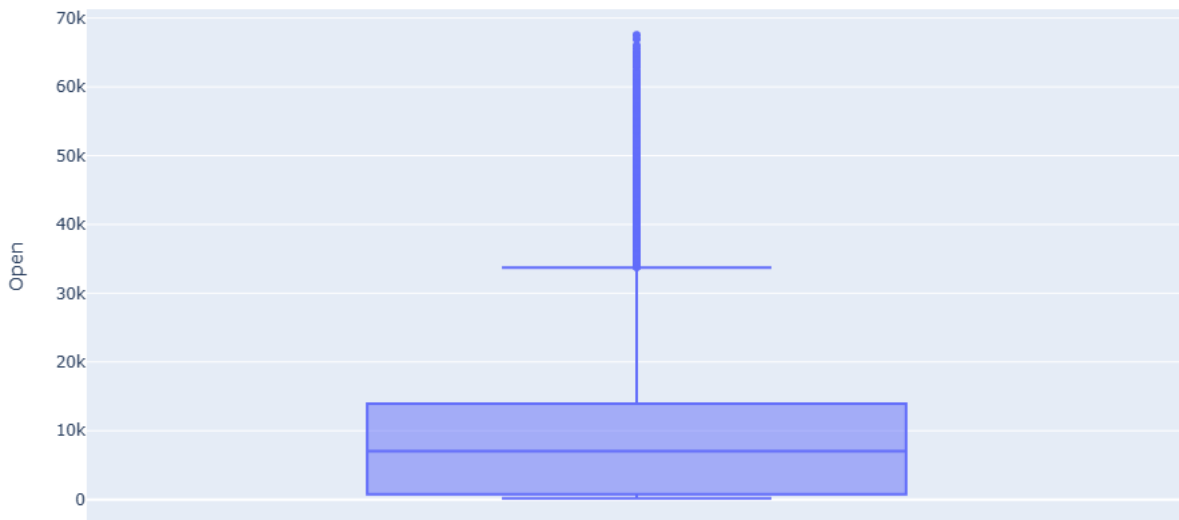


Figure 3.11: Box Plot of Open

### 3.4 Model Specification

Our research deal with a time series problem where we will forecast the price of the bitcoin into the future with higher accuracy using an Ensemble Model built with Bi\_LSTM and Random Forest Regression considering macro economic and media influence factors and create a comparison of our proposed model with a regular Bi\_LSTM with all the selected features. Furthermore, we have analyzed the other time series models and generated a comparative study with models such as FB Prophet, LSTM, GRU and also Bidirectional GRU for our result analysis purpose.

### 3.4.1 Vader (Valence Aware Dictionary sEntiment Reasoning)

The Vader model was used to calculate the polarity from the tweets that we have collected. Vader is a trained model that can classify the sentences into categories. We have classified the scores into 4 types(positive, negative, neutral and compound). Vader does not require pre-processing such as tokenization and lemmatization is not required. Punctuations, capitalization and special characters such as hashtags and emojis can also be identified.

To calculate the compound score, Vader uses the following formula:

$$compound = (P - N)/(P + N + N_C) \quad (3.2)$$

where: P is the addition of the positive scores of all words in the text

N is the addition of the negative scores of all words in the text

N\_C is the addition of the neutral scores of all words in the text

The compound score is a normalized score between -1(extreme negative) to +1(extreme positive).

### 3.4.2 LSTM (Long Short Term Memory)

The vanishing gradient problem is addressed by LSTM. The LSTM model uses an input gate, forget gate and output gate for sustaining the data that is relevant for making the forecast. LSTM remembers the information which is important for forecasting. The following equations related to the LSTM model are:

$$InputGate = (w_i * x_t + u_i * h_{t-1} + b_i) \quad (3.3)$$

$$ForgetGate = (w_f * x_t + u_f * h_{t-1} + b_f) \quad (3.4)$$

$$OutputGate = (w_o * x_t + u_o * h_{t-1} + b_o) \quad (3.5)$$

$$MemoryCell = f_t * c_{t-1} + i_t * \tanh(w_c * x_t + u_c * h_{t-1} + b_c) \quad (3.6)$$

$$CurrentHiddenState = o_t * \tanh(c_t) \quad (3.7)$$

where,

$x_t$  is the input at time step

$t$ ,  $h_{t-1}$  is the previous hidden state

$c_{t-1}$  is the previous memory cell state

$w_i, u_i, b_i, w_f, u_f, b_f, w_o, u_o, b_o, w_c, u_c, b_c$  are the weights and biases of the gates and memory cell.

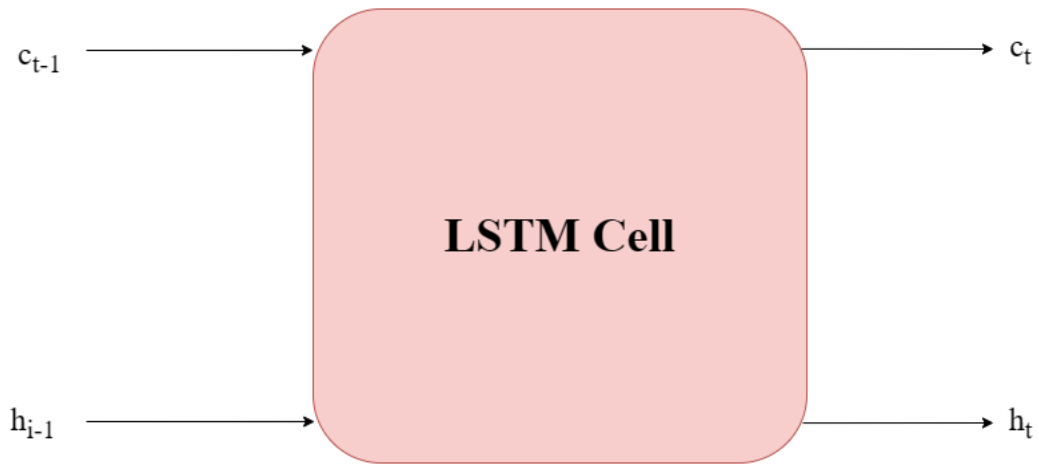


Figure 3.12: LSTM Cell

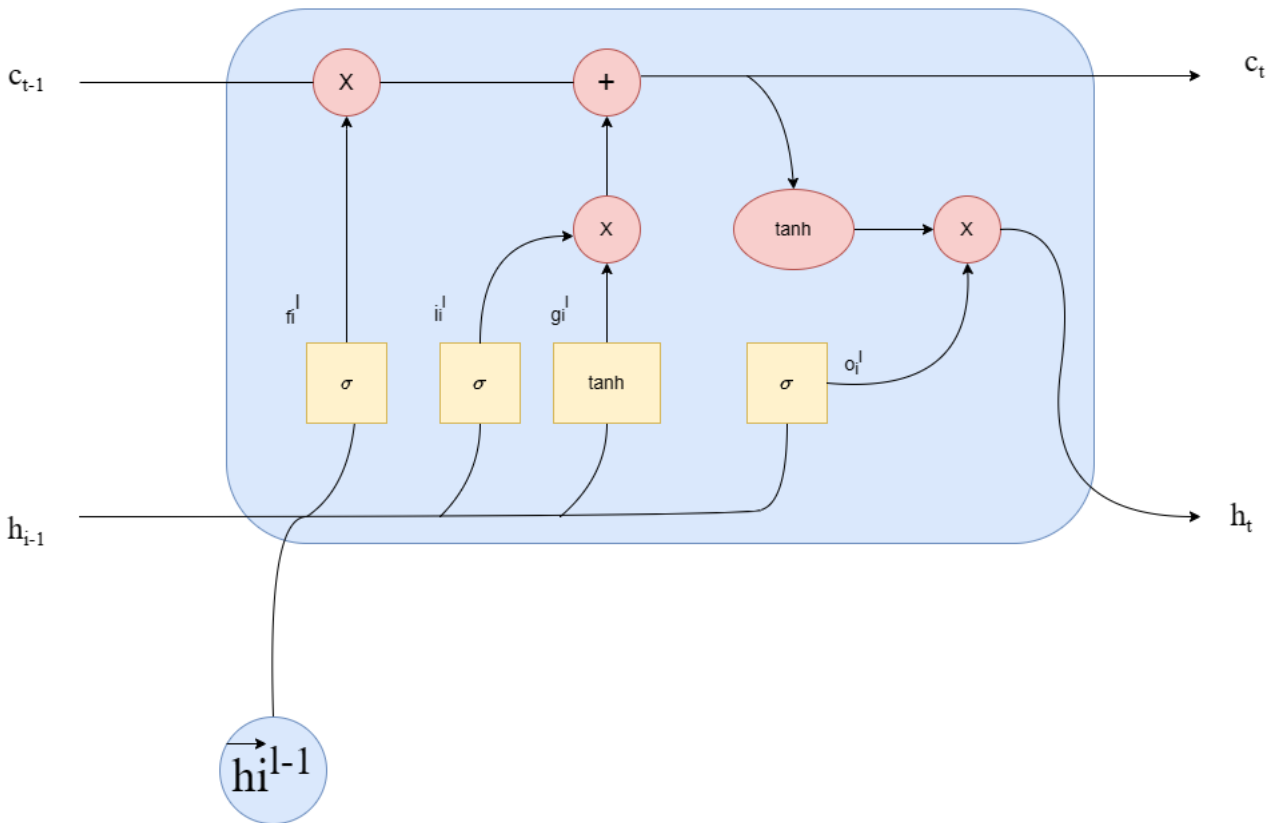


Figure 3.13: LSTM Cell Details

### 3.4.3 GRU (Gated Recurrent Unit)

GRU is similar to LSTM, but it has fewer parameters and a more simpler structure. GRU has two gates; the first one is the update gate and the second one is the reset gate. The update gate regulates the hidden states to be kept and how much of the hidden states should be removed. The formula of update gate is:

$$UpdateGate = (w_z * x + u_z * h_{previous} + b_z) \quad (3.8)$$

Where,

$x$  = current input

$h_{previous}$  = previous hidden state

$w_z, u_z$  and  $b_z$  = weights and biases of the update gate

The reset gate decides how much of the previous data should be forgotten. The reset gate formula is:

$$ResetGate = (w_r * x + u_r * h_{previous} + b_r) \quad (3.9)$$

Where,

$x$  = current input

$h_{previous}$  = previous hidden state

$w_r, u_r$  and  $b_r$  = weights and biases of the reset gate

Candidate Hidden State are the temporal hidden states where the values would later be updated to the current hidden state. The formula of candidate hidden state is:

$$CandidateHiddenState = \tanh(W_h * x + U_h * (r * h_{prev}) + b_h) \quad (3.10)$$

### 3.4.4 Bidirectional GRU

Bidirectional GRU has the same structure as GRU but it processes the data in backward and forward direction. This allows the model to work on both the past and future context when making decisions. Here, the model contains two layers. One layer works with the input sequence in the front direction whereas the other layer operates in the backward direction. The forward pass uses the following formulas:

$$Updategate(z_f) : z_f = (W_{zf} * x + U_{zf} * h_{previous} + b_{zf}) \quad (3.11)$$

$$Resetgate(r_f) : r_f = (W_{rf} * x + U_{rf} * h_{previous} + b_{rf}) \quad (3.12)$$

$$Candidatehiddenstate(h_{cf}) : h_{cf} = \tanh(W_{hf} * x + U_{hf} * (r_f * h_{previous}) + b_{hf}) \quad (3.13)$$

$$Currenthiddenstate(h_{tf}) : h_{tf} = (1 - z_f) * h_{previous} + z_f * h_{cf} \quad (3.14)$$

The backward pass uses the following formulas:

$$Updategate(z_b) : z_b = (W_{zb} * x + U_{zb} * h_{previous} + b_{zb}) \quad (3.15)$$

$$Resetgate(r_b) : r_b = (W_{rb} * x + U_{rb} * h_{previous} + b_{rb}) \quad (3.16)$$

$$\text{Candidatehiddenstate}(h_{cb}) : h_{cb} = \tanh(W_{hb} * x + U_{hb} * (r_b * h_{previous}) + b_{hb}) \quad (3.17)$$

$$\text{Currenthiddenstate}(h_{tb}) : h_{tb} = (1 - z_b) * h_{previous} + z_b * h_{cb} \quad (3.18)$$

Where  $x$  is the current input,  $h_{previous}$  is the previous hidden state,  $W_z f$ ,  $U_z f$ ,  $b_z f$ ,  $W_r f$ ,  $U_r f$ ,  $b_r f$ ,  $W_h f$ ,  $U_h f$ ,  $b_h f$ ,  $W_z b$ ,  $U_z b$ ,  $b_z b$ ,  $W_r b$ ,  $U_r b$ ,  $b_r b$ ,  $W_h b$ ,  $U_h b$ ,  $b_h b$  are the weights and biases of the forward and backward gates respectively.

### 3.4.5 FB Prophet

FBProphet is an open-source time series forecasting library designed explicitly for forecasting at scale. It is built on decomposing time series data into its three main components: trend, seasonality, and holidays. Prophet uses a time series model with three main model components:

1. a piecewise linear curve.
2. Fourier series to analyze the annual seasonality component.
3. Weekly seasonal component through the use of dummy variables

It also can include multiple change points to allow for non-linear trends in the data. The Prophet model is mathematically based on the additive decomposition of a time series  $y(t)$  into three components: trend  $f(t)$ , seasonality  $s(t)$ , and holidays  $h(t)$ , as well as an error term  $e(t)$  :

$$y(t) = f(t) + s(t) + h(t) + e(t) \quad (3.19)$$

The trend component  $f(t)$  is modeled using a piecewise linear function or a logistic function, depending on whether the forecast is for a continuous or binary variable. The seasonality component  $s(t)$  is modeled using the Fourier series, and the holidays component  $h(t)$  is modeled using a set of binary variables. Prophet also allows for user-specified regressors, which can be used to incorporate external variables like price and weather and can be modified to make multivariate analyses.

### 3.4.6 Bidirectional LSTM

Bidirectional LSTMs are an extension of regular LSTMs that can enhance the performance of models for sequence classification and time series challenges. Bidirectional LSTMs train two instead of one LSTM on the input sequence when all timesteps of the input sequence are accessible. The first occurs on the original input sequence, while the second occurs on a reversed copy. This can provide more context to the network, resulting in faster and more comprehensive problem-solving learning. The functions at the gates of basic Bidirectional LSTM cell units are the following:

$$\text{ForgetGate} : f_t = (W_f * [h_{t-1}, x_t] + b_f) \quad (3.20)$$

$$\text{InputGate} : i_t = (W_i * [h_{t-1}, x_t] + b_i) \quad (3.21)$$

$$\text{OutputGate} : o_t = (W_o * [h_{t-1}, x_t] + b_o) \quad (3.22)$$

## Our Proposed Bi-LSTM Model for Research

For training our model, we used 4 hidden layers, each with 100 neurons and a dropout rate of 0.2 with a batch size of 100. The sigmoid function is our activation function, the loss function is the mean square error, and optimizers are used as adam. Later, to encompass all the macroeconomic factors and twitter's influence on the media, we trained four distinct Bi-LSTM on the following datasets. Which are:

1. Model 1: Bidirectional LSTM, Input Feature: Open, High, Low, Volume
2. Model 2: Bidirectional LSTM with Sentiment Analysis, Input Feature: Open, High, Low, Volume, Tweet Count, Avg Compound Score, Min Compound Score, Max Compound Score
3. Model 3: Bidirectional LSTM with AMD & NVIDIA Stock Price Input Feature: Open, High, Low, Volume, AMD\_Open, NVIDIA\_Open
4. Model 4: Bidirectional LSTM with Economic Factors Input Feature: Open, High, Low, Volume, Gold\_Open, S&P500\_Open, NASDAQ\_Open

Then taking prediction and forecast from all four models, we combined them into a single dataset one for prediction and one for forecast for feeding into the ensemble model implemented with Random Forest Regressor. From the final ensemble model our target is to forecast the Open price of Bitcoin.

### 3.4.7 Random Forest

The basic idea behind Random Forest is to construct many decision trees and then average their predictions. The model works by repeatedly splitting the data into subsets based on the values of certain features. Each split results in a new "node" in the decision tree, and each leaf node represents a predicted value for the target variable.

Random Forest is an ensemble method that combines the predictions of multiple decision trees to improve the model's overall performance. So in our case, we combined the prediction of multiple Bidirectional LSTM models and applied natural logarithmic transformation, then used the transformed dataset for training the Random Forest model itself. Later we used the trained model to make the forecast by using the data of the forecast we got from the four Bidirectional LSTM models. This way, we improved the accuracy of our prediction. Also, we solved the memoryless problem of Random Forest by using the forecast of the Bi-LSTMs model but still having the accuracy intact.

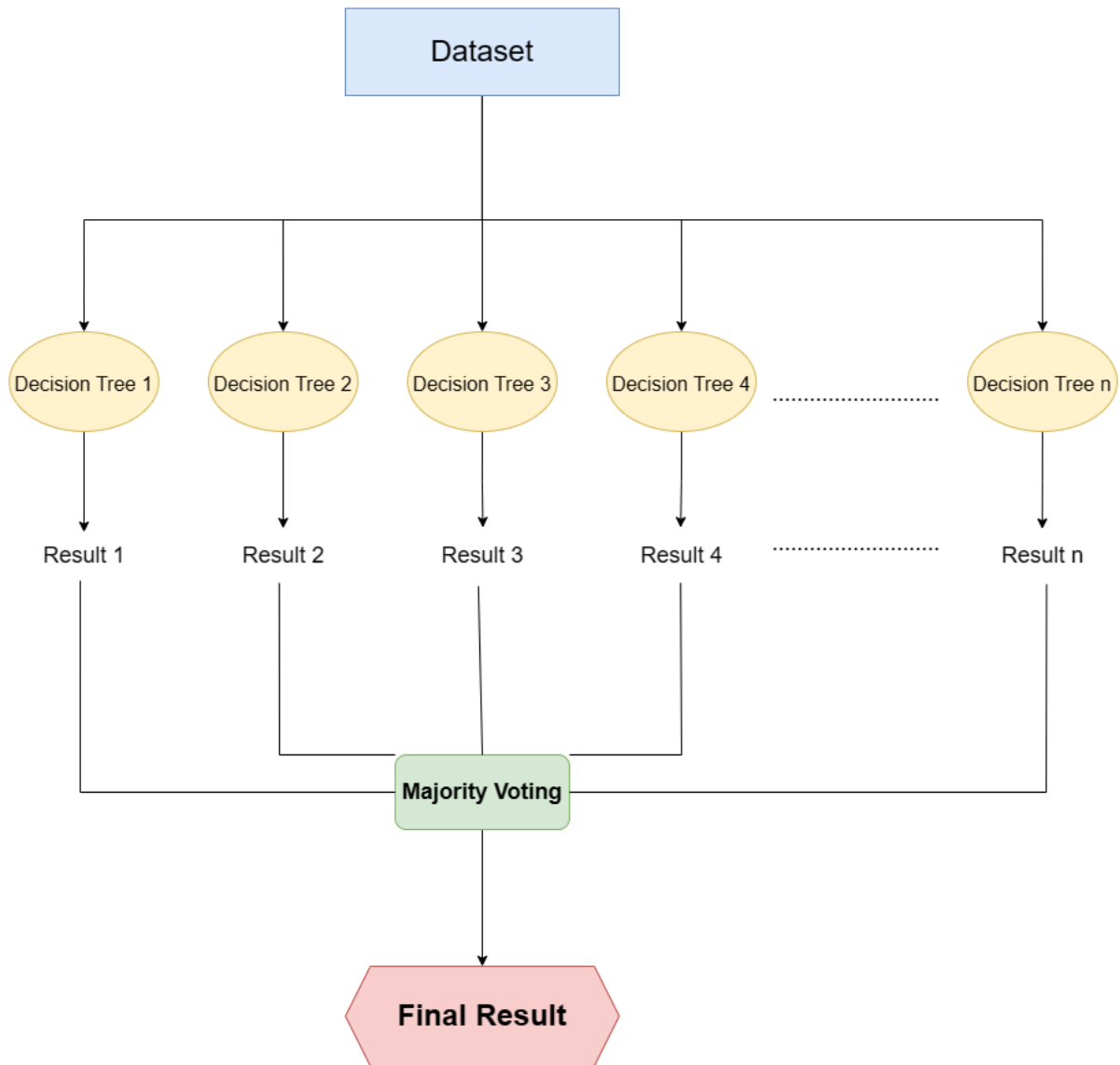


Figure 3.14: Random Forest

### 3.4.8 Overall Model Specification

Our research deal with a time series problem where we will forecast the price of the bitcoin into the future with higher accuracy using an Ensemble Model built with Bi\_LSTM and Random Forest Regression considering macro economic and media influence factors and create a comparison of our proposed model with a regular Bi\_LSTM with all the selected features. Furthermore, we have analyzed the other time series models and generated a comparative study with models such as FB Prophet, LSTM, GRU and also Bidirectional GRU for our result analysis purpose. The Image below shows our model specification diagram to our proposed model:



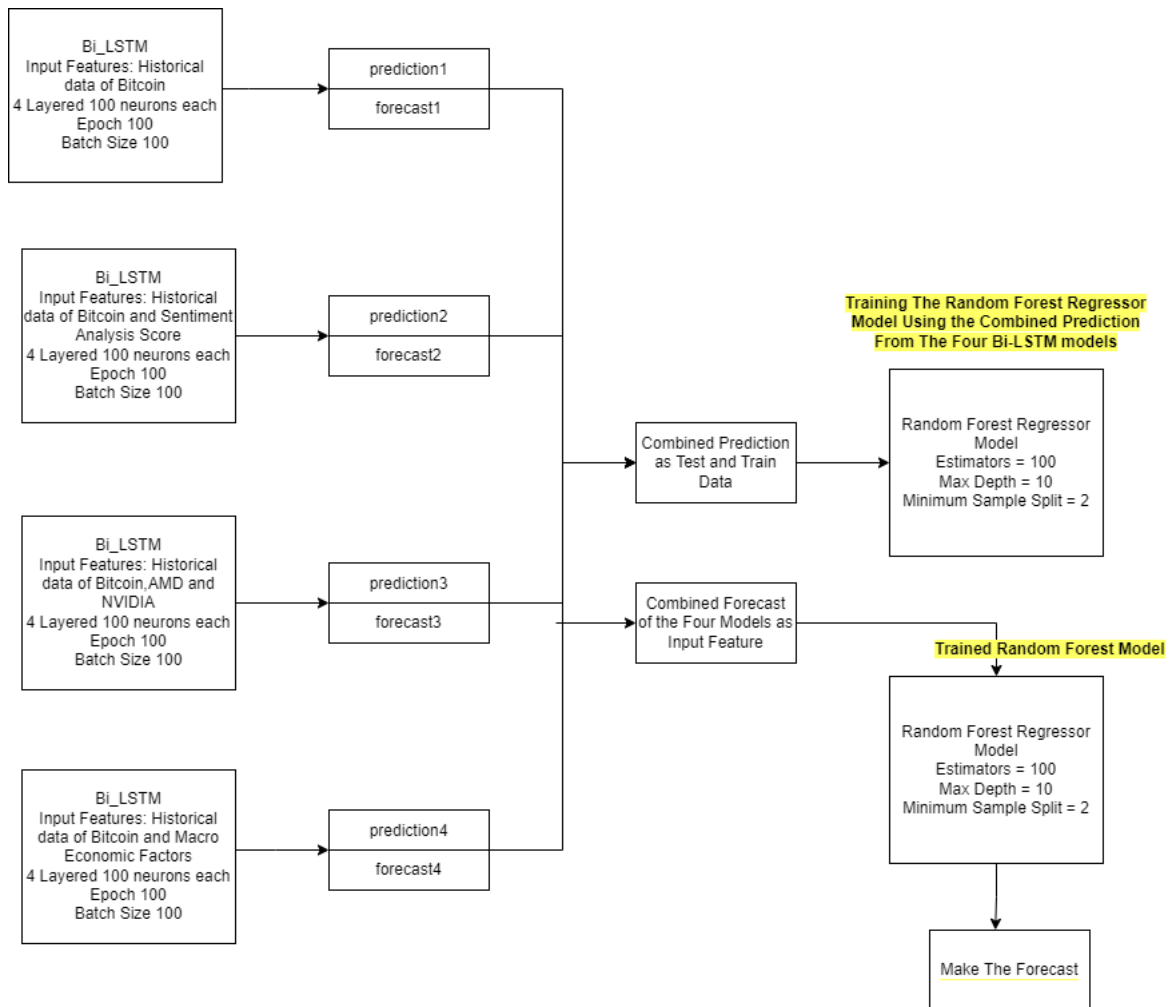


Figure 3.15: Overall Model Specification

# Chapter 4

## Results and Discussion

### 4.1 Error Matrices

From our proposed models, we have generated the results that provide us with significant insights about the trends and variations in the price of Bitcoin that results in huge volatility. In our research we evaluated each model with MSE, MAE, R2 and MAPE performance scores. Furthermore, we ran each model 20 times to find the deviation in the performance scores.

1. Mean Squared Error (MSE) is a commonly used metric for evaluating the accuracy of time series forecasts (Lower is Better). It is the average of the squared differences between the predicted values and the actual values. The formula for MSE is:

$$MSE = (1/n) * \sum (y_i - y^i)^2 \quad (4.1)$$

where  $y_i$  is the actual value,  $y^i$  is the predicted value, and n is the number of observations.

2. Mean Absolute Error (MAE) is another metric for evaluating forecast accuracy (Lower is better). It is the average of the absolute differences between the predicted values and the actual values. The formula for MAE is:

$$MAE = (1/n) * |y_i - y^i| \quad (4.2)$$

3. R-squared (R2) is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). In the case of time series forecasting (Higher is better), it represents the proportion of the variance in the actual values that is predictable from the predicted values. The formula for R2 is:

$$R2 = 1 - (SS_{residual}/SS_{total}) \quad (4.3)$$

where  $SS_{residual}$  is the sum of squares of the residuals (predicted - actual) and  $SS_{total}$  is the total sum of squares (actual - mean of actual).

4. Mean Absolute Percentage Error (MAPE) is a measure of forecast accuracy of a model (Lower is better). It calculates the average absolute error as a percentage of the actual values. The formula for MAPE is:

$$MAPE = (100/n) * |(y_i - y^i)/y_i| \quad (4.4)$$

## 4.2 Result Analysis

In our research, we have selected Bi\_LSTM as our base neural network model to create our ensemble model. Therefore, we have drawn a comparative analysis with some other significant multivariate time series models and further proved that Bi\_LSTM performs better in predicting the Bitcoin price. We have used FBProphet, GRU, Bi\_GRU and LSTM to show the comparison with Bi\_LSTM. The table below shows the comparative analysis of the models.

	<b>MSE</b>	<b>MAE</b>	<b>R2</b>	<b>MAPE</b>	<b>p-value</b>
<b>Bi LSTM</b>	827.9759	501.2073	0.997593267	0.104931601	0.01205
<b>FB Prophet</b>	82.6563	68.3486	0.5431	0.2186	0.0004337
<b>GRU</b>	1094.4535	748.3706	0.99579	0.144208	0.001062
<b>Bi GRU</b>	1262.3828	821.0536	0.994405	0.137284	0.001050
<b>LSTM</b>	1127.7614	675.0086	0.99553	0.2221739	0.00117

Table 4.1: Comparative Analysis of Multiple Time Series Models

Each of the selected models does not require the data to be stationary as we have not transformed our original data. Here, from the table, we can see that each of the models has a p-value from t-test is less than 0.05 (5%) which proves against the null hypothesis that each of the models is statistically significant in predicting the Bitcoin price and the result produced is not just random. Although FBProphet shows good performance in MSE and MAE, the R2 score is significantly lower than the other models. Bi\_LSTM on the other hand shows good performance in all the matrices. Another well known model which is ARIMA is not present in our analysis as previous studies [10] have already proved that other models perform better than ARIMA.

Following our first analysis, we trained our four parallel models and compared them with the selected 4 performance metrics. The following table [4.2] shows the performance comparison of the 4 models.

From the table [4.2], we can see that Bi\_LSTM with Sentiment Analysis gives the best R2 score of 0.9976736092963 with a deviation of +- 0.000359, MAE score of 518.81492841468 with a deviation of +- 43.3999, MSE score of 814.03877324874 with a deviation of +- 57.8609 and best MAPE score of 0.1233077542727 with a deviation of +- 0.02511. Bi\_LSTM with Sentiment Analysis gives a better score, but Bi\_LSTM shows least deviation in MAPE and MAE while Bi\_LSTM with Economic Factors shows least deviation in MSE and R2. The bar chart in figure [4.1]

	MSE	MAE	R2	MAPE
<b>Bi_LSTM</b>	1120.375415 +- 51.8746	725.13471298 +- 30.9969	0.9955932378 +- 0.00034	0.174953240 +- 0.0151
<b>Bi_LSTM with Sentiment Analysis</b>	814.0387732 4874 +- 57.8609	518.814928414 68 +- 43.3999	0.99767360929 63 +- 0.000359	0.123307754 2727 +- 0.02511
<b>Bi_LSTM with AMD,NVIDIA</b>	877.149024 +- 136.8191	597.092425 +- 71.3611	0.99729890 +- 0.001017	0.18823974 +- 0.02612
<b>Bi_LSTM with Economic Factors</b>	918.1968982 +- 50.0208	625.3410641 +- 34.8834	0.99704018 +- 0.000318	0.14316915 +- 0.04799

Table 4.2: Performance score of the four models

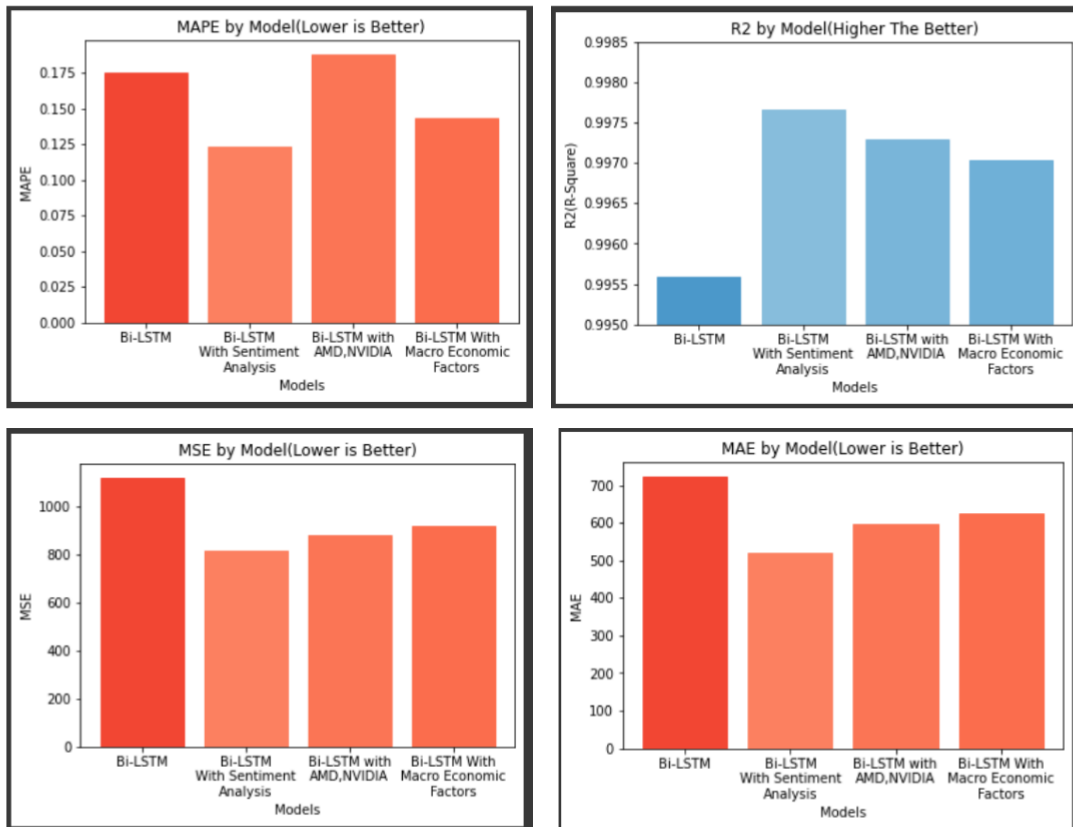


Figure 4.1: Performance score of the four models

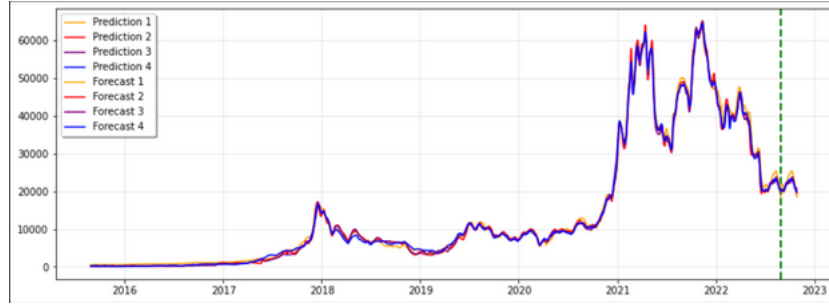


Figure 4.2: Prediction of all four models

below visualizes the performance scores of the 4 models. In the ensemble model, the RFR was trained using the combined prediction so that it can select an accurate forecast from the combined forecast we got from the above 4 models. The Table below shows the performance score of our ensemble model on the test size of the predictions of the 4 previous models as input compared with the Bi\_LSTM model with all the features from the previous 4 models as input.

Here we can see that the performance of our ensemble model performs significantly

	MSE	MAE	R2	MAPE
<b>Random Forest Ensemble Model</b>	0.0021607409 +-0.000211	0.031870944 +-0.00092	0.9990956883 +-8.8897e-5	0.0038217062 +-0.000126
<b>Bi_LSTM Macro Economic Factors and Sentiment Analysis</b>	827.9759195 +- 49.9204	501.2073303 +- 33.5967	0.997593267 +- 0.0003205	0.104931601 +- 0.024238

Table 4.3: Prediction of all 4 models

well in all of the performance matrices with significant small deviation in each score. Also our ensemble model outperforms the other 4 Bi\_LSTM model which clearly shows that the ensemble model is selecting the best from the 4 models. Finally, we make a forecast of 60 days into the future using our ensemble model as it is the only model that performs significantly better than any other models from our study. The following graph shows the forecast form our ensemble model.

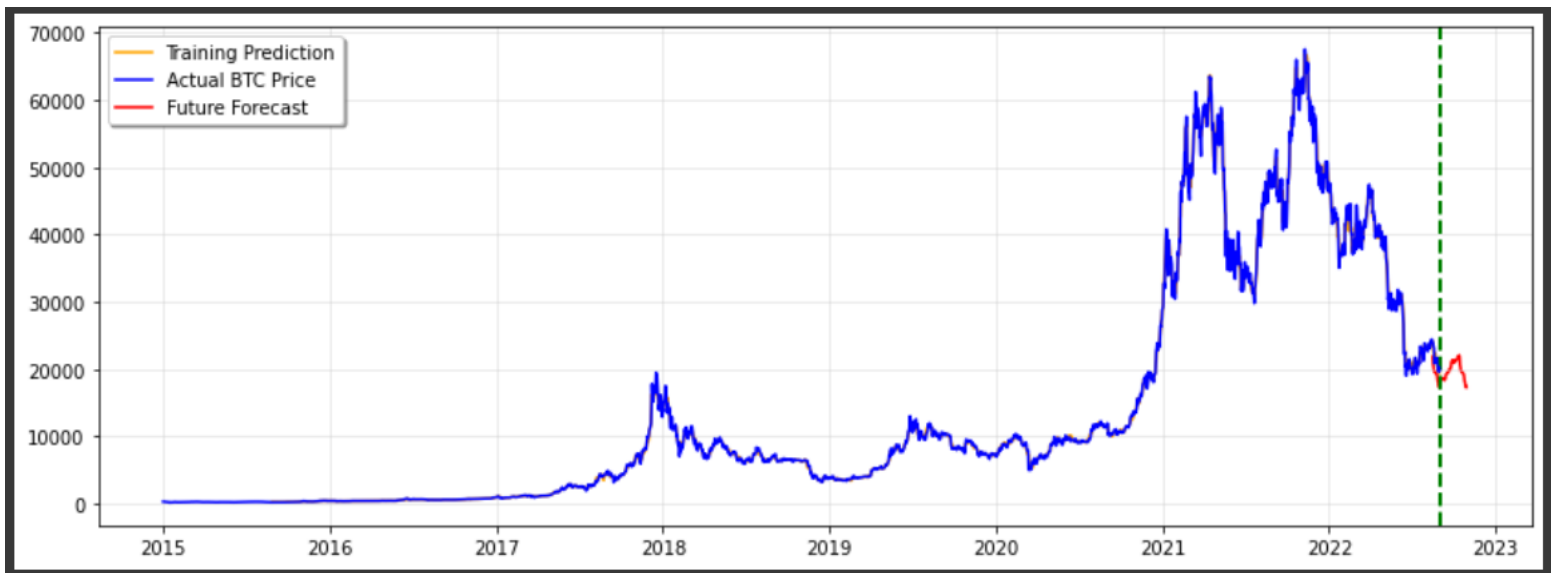


Figure 4.3: Final Forecast Graph for Ensemble Model

# Chapter 5

## Conclusion and Future Work

To sum up, we have seen that our Bi-LSTM with ensemble Random Forest Regressor Model has performed better. The result parameters MSE, MAE, R2 and MAPE have outperformed the combined Bi-LSTM model with Macro Economic factors and Sentiment Analysis. Moreover, we have proved that Bi-LSTM with an ensemble Supervised Random Forest Regression Model will give more accurate predictions. For our future work, we would be implementing Contextual Data, like Geopolitical Data such as War, political movement, Economic Depression or Famine. We would like to generate an error band to estimate the high and low values of our prediction.

# Chapter 6

## References

- [1] Yiyiing, W., Yeze, Z. (2019). Cryptocurrency Price Analysis with Artificial Intelligence. IEEE Xplore. <https://doi.org/10.1109/INFOMAN.2019.8714700>
- [2] Kalariya, V., Parmar, P., Tanwar, S., Kumar, N., Alazab, M. (2020). Stochastic Neural Networks for Cryptocurrency Price Prediction. IEEE Access, 8, 82804–82818. <https://doi.org/10.1109/access.2020.299065>
- [3] Dipple, S., Choudhary, A., Flamino, J., Szymanski, B. K., Korniss, G. (2020). Using correlated stochastic differential equations to forecast cryptocurrency rates and social media activities. Applied Network Science. <https://doi.org/10.1007/s41109-020-00259-1>
- [4] Yang Li, Zibin Zheng and Hong-Ning Dai (2020). Enhancing Bitcoin Price Fluctuation Prediction Using Attentive LSTM and Embedding Network. 19 Appl. Sci. 2020, 10(14), 4872. <https://doi.org/10.3390/app10144872>
- [5] Tandon, Sakshi, et al. “Bitcoin Price Forecasting Using LSTM and 10- Fold Cross Validation.” IEEE Xplore, 1 Mar. 2019, [ieeexplore.ieee.org/document/8938251](https://ieeexplore.ieee.org/document/8938251). <https://doi.org/10.1109/ICSC45622.2019.8938251>
- [6] Livieris, Ioannis E., et al. “An Advanced CNN-LSTM Model for Cryptocurrency Forecasting.” Electronics, vol. 10, no. 3, 26 Jan. 2021, p. 287, [10.3390/electronics10030287](https://doi.org/10.3390/electronics10030287). Accessed 13 Mar. 2021. <https://doi.org/10.3390/electronics10030287>
- [7] Kim, Jong-Min, et al. “Forecasting the Price of the Cryptocurrency Using Linear and Nonlinear Error Correction Model.” Journal of Risk and Financial Management, vol. 15, no. 2, 10 Feb. 2022, p. 74, [10.3390/jrfm15020074](https://doi.org/10.3390/jrfm15020074). Accessed 1 Apr. 2022. <https://doi.org/10.3390/jrfm15020074>
- [8] Hamayel, M. J., Owda, A. Y. (2021). A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms. AI, 2(4), 477–496. IEEE Standard for General Requirements for Cryptocurrency Exchanges. (2020). IEEE Std 2140.1-2020, 1–18. <https://doi.org/10.1109/IEEESTD.2020.9248667>
- [9] Sun, J., Zhou, Y., Lin, J. (n.d.). Using machine learning for cryptocurrency



trading.

- [10] Hashish, I., Forni, F., Andreotti, G., Facchinetti, T., Darjani, S. (n.d.). A Hybrid Model for Bitcoin Prices Prediction using Hidden Markov Models and Optimized LSTM Networks.
- [11] Dao, Nguyen-An, et al. “Predicting Cryptocurrency Price Movements Based on Social Media.” 2019 International Conference on Advanced Computing and Applications (ACOMP), Nov. 2019, 10.1109/acomp.2019.00016. Accessed 16 Sep. 2021.
- [12] [2204.05783] Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets. (2022, April 7). arXiv. Retrieved January 10, 2023, from <https://arxiv.org/abs/2204.05783>
- [13] Inamdar, Abid, et al. “Predicting Cryptocurrency Value Using Sentiment Analysis.” IEEE Xplore, 1 May 2019, [ieeexplore.ieee.org/abstract/document/9065838](http://ieeexplore.ieee.org/abstract/document/9065838).
- [14] Serafini, Giulia, et al. “Sentiment-Driven Price Prediction of the Bitcoin Based on Statistical and Deep Learning Approaches.” IEEE Xplore, 1 July 2020, [ieeexplore.ieee.org/document/9206704/authorsauthors](http://ieeexplore.ieee.org/document/9206704/authorsauthors).
- [15] Cakra, Yahya Eru, and Bayu Distiawan Trisedya. “Stock Price Prediction Using Linear Regression Based on Sentiment Analysis.” 2015 International Conference on Advanced Computer Science and Information Systems (ICACISIS), vol. 15804429, Oct. 2015, 10.1109/icacsis.2015.7415179.
- [16] Gao, Tingwei, et al. “Applying Long Short Term Memory Neural Networks for Predicting Stock Closing Price.” 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Nov. 2017, 10.1109/icseess.2017.8342981.
- [17] Gupta, Aditya, and Bhuwan Dhingra. “Stock Market Prediction Using Hidden Markov Models.” 2012 Students Conference on Engineering and Systems, Mar. 2012, [ieeexplore.ieee.org/document/6199099/](http://ieeexplore.ieee.org/document/6199099/), 10.1109/sces.2012.6199099.
- [18] Hariadi, Mochammad, et al. “Prediction of Stock Prices Using Markov Chain Monte Carlo.” IEEE Xplore, 1 Nov. 2020, [ieeexplore.ieee.org/document/9297965](http://ieeexplore.ieee.org/document/9297965). Accessed 11 Jan. 2023.
- [19] Fu, Rui, et al. “Using LSTM and GRU Neural Network Methods for Traffic Flow Prediction.” 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Nov. 2016, 10.1109/yac.2016.7804912.
- [20] Pan, Mingyang, et al. “Water Level Prediction Model Based on GRU and CNN.” IEEE Access, vol. 8, 2020, pp. 60090–60100, 10.1109/access.2020.2982433. Accessed 4 Dec. 2020.
- [21] Zhai, Naiju, et al. “Multivariate Time Series Forecast in Industrial Process Based on XGBoost and GRU.” IEEE Xplore, 1 Dec. 2020, [ieeexplore.ieee.org/document/9338878](http://ieeexplore.ieee.org/document/9338878).

- [22] Wenjie Lu, et al. A CNN-LSTM-Based Model to Forecast Stock Prices. School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, 22 Oct. 2020.
- [23] Rizwan, Muhammad, et al. “Bitcoin Price Prediction Using Deep Learning Algorithm.” IEEE Xplore, 1 Dec. 2019, [ieeexplore.ieee.org/document/9024772](https://ieeexplore.ieee.org/document/9024772).
- [24] Chen, S., Zhou, C. (2021). Stock Prediction Based on Genetic Algorithm Feature Selection and Long Short-Term Memory Neural Network. IEEE Access, 9, 9066–9072. <https://doi.org/10.1109/access.2020.3047109>
- [25] Gorgolis, N., Istenes, Z., Gyenne, L.-G. (n.d.). Hyperparameter Optimization of LSTM Network Models through Genetic Algorithm.
- [26] Mootha, S., Sridhar, S., Seetharaman, R., Chitrakala, S. (n.d.). Stock Price Prediction using Bi-Directional LSTM based Sequence to Sequence Modeling and Multitask Learning.
- [27] Pandey, A., Misra, S., Saxena, M. (n.d.). Gold and Diamond Price Prediction Using Enhanced Ensemble Learning.
- [28] He, Z., Zhou, J., Dai, H.-N., Wang, H. (2019). Gold Price Forecast Based on LSTM-CNN Model. 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCOM/CyberSciTech) . <https://doi.org/10.1109/dasc/picom/cbdcom/cyberscitech.2019.00188>
- [29] S, V. G., S, H. V. (2020). Gold Price Prediction and Modelling using Deep Learning Techniques. 2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS). <https://doi.org/10.1109/raics51191.2020.9332471>
- [30] Gabralla, L., Jammazi, R., Abraham, A. (n.d.). 2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE) 674 Oil Price Prediction Using Ensemble Machine Learning.
- [31] Shang, Qiang, et al. “A Hybrid Method for Traffic Incident Detection Using Random Forest-Recursive Feature Elimination and Long Short-Term Memory Network with Bayesian Optimization Algorithm.” IEEE Access, vol. 9, 2021, pp. 1219–1232, [ieeexplore.ieee.org/document/9306801](https://ieeexplore.ieee.org/document/9306801), 10.1109/ACCESS.2020.3047340. Accessed 11 Jan. 2023.
- [32] Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., Lama, B. K. (2018). Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis. 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS). <https://doi.org/10.1109/cccs.2018.8586824>

- [33] Dipple, S., Choudhary, A., Flamino, J., Szymanski, B. K., Korniss, G. (2020). Using correlated stochastic differential equations to forecast cryptocurrency rates and social media activities. *Applied Network Science*, 5(1). <https://doi.org/10.1007/s41109-020-00259-1>
- [34] Rathan, K., Sai, S. V., Manikanta, T. S. (2019). Crypto-Currency price prediction using Decision Tree and Regression techniques. *IEEE Xplore*. <https://doi.org/10.1109/ICOEI.2019.8862585>
- [35] Phaladisailoed, T., Numnonda, T. (2018). Machine Learning Models Comparison for Bitcoin Price Prediction. 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE). <https://doi.org/10.1109/iciteed.2018.8534911>
- [36] Felizardo, L., Oliveira, R., Del-Moral-Hernandez, E., Cozman, F. (2019, October 1). Comparative study of Bitcoin price prediction using WaveNets, Recurrent Neural Networks and other Machine Learning Methods. *IEEE Xplore*. <https://doi.org/10.1109/BESC48373.2019.8963009>
- [37] Jones, E. (2022, January 7). A Brief History of Cryptocurrency. *CryptoVantage*. <https://www.cryptovantage.com/guides/a-brief-history-of-cryptocurrency/>
- [38] Kasra, S., Jönköping, Shariati, K., Sjölander, P. (2022). The Influence of Gold Market on Bitcoin Prices Business Administration -Finance NUMBER OF CREDITS: 15 ECTS PROGRAMME OF STUDY: International Financial Analysis Is there a correlation? i Title: The Influence of Gold Market on Bitcoin Prices. <https://www.diva-portal.org/smash/get/diva2:1672975/FULLTEXT01.pdf>
- [39] Kumah, S. P., Odei-Mensah, J. (2022). Do cryptocurrencies and crude oil influence each other? Evidence from wavelet-based quantile-in-quantile approach. *Cogent Economics Finance*, 10(1). <https://doi.org/10.1080/23322039.2022.2082027>
- [40] Reiff, N. (2020, June 22). Should You Buy Gold Or Bitcoin? *Investopedia*. <https://www.investopedia.com/news/should-you-buy-gold-or-bitcoin/>
- [41] Godbole, Omkar. “Bitcoin’s Correlation to SP 500 Hits 17-Month High.” *Www.coindesk.com*, 23 Mar. 2022, [www.coindesk.com/markets/2022/03/23/bitcoins-correlation-to-sp-500-hits-17-month-high/](http://www.coindesk.com/markets/2022/03/23/bitcoins-correlation-to-sp-500-hits-17-month-high/).
- [42] DeMatteo, Megan. “Bitcoin Is on Its Way to \$100,000, According to Experts. Here’s When They Predict It Will Happen.” *Time*, 22 Nov. 2021, [time.com/nextadvisor/investing/crypto-price-predictions/](http://time.com/nextadvisor/investing/crypto-price-predictions/).
- [43] “Bitcoin USD (BTC-USD) Stock Historical Prices Data.” @YahooFinance, 2010, [finance.yahoo.com/quote/BTC-USD/history?p=BTC-USD](http://finance.yahoo.com/quote/BTC-USD/history?p=BTC-USD).
- [44] “The Role of Social Media in Crypto Prices — Mawson.” *Mawson Infrastructure Group*, 4 Feb. 2022, [mawsoninc.com/role-of-social-media-in-crypto-prices/](http://mawsoninc.com/role-of-social-media-in-crypto-prices/).

[45] Aaron Klotz. “GPU Prices Drop along with Crypto.” Tom’s Hardware, 24 Jan. 2022, [www.tomshardware.com/news/gpu-prices-plummet-along-with-crypto](http://www.tomshardware.com/news/gpu-prices-plummet-along-with-crypto).