Multi-Head Self-Attention Transformer for Dogecoin Price Prediction

Sashank Sridhar
Department of Computer Science and Engineering
College of Engineering Guindy, Anna University
Chennai, India
sashank.ssridhar@gmail.com

Sowmya Sanagavarapu
Department of Computer Science and Engineering,
College of Engineering Guindy, Anna University
Chennai, India
sowmya.ssanagavarapu@gmail.com

2021 14th International Conference on Human System Interaction (HSI) | 978-1-6654-4112-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/HSI52170.2021.9538640

***Abstract*— Cryptocurrency market has witnessed a boom during the global pandemic and has proven as a strong investment with a wide institutional adoption. A time-series forecasting solution will play a vital role in analyzing the fluctuation of the bitcoin and altcoin markets. Dogecoin is one such altcoin that is a low-price, high-risk investment option garnering considerable interest this year. The variation of the price trend of this altcoin is studied using the multi-head attention mechanism implemented in a transformer, where the attention heads attend to the tokens that are relevant to each current token based on varying short-term and long-term dependencies. In this paper, a multi-head attention-based transformer encoder-decoder model is applied on the hourly data of the Dogecoin price for its prediction over time. The performance of the model has been evaluated using a number of evaluation metrics including MAE and predictive R-squared value. The model trained over the Dogecoin hourly price variation gave an impressive accuracy of 98.46% and R-squared value of 0.8616 comparable with the existing state-of-the-art cryptocurrency price forecasting models.**

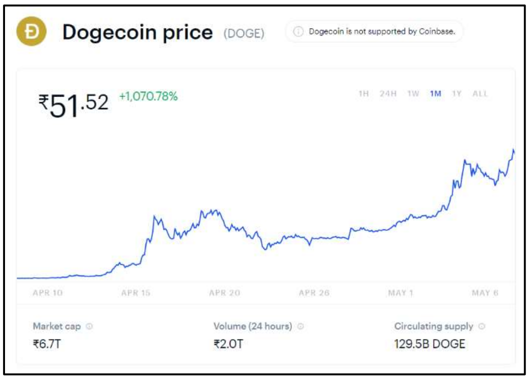
Keywords*— Transformer, Self-Attention, Multi-head Attention, Cryptocurrency Price Prediction

I. Introduction

Cryptocurrencies aim to provide user autonomy over their money more than fiat currencies along with low transaction fee and no banking fee [1]. The high accessibility [2] of these currencies without disclosure of personal information and traceability, enabled their growth over the years. Digital assets such as tokens and cryptocurrencies excluding bitcoin are known as alternative cryptocurrencies, or altcoins [3] in short. Dogecoin [4] is one such altcoin that has gained traction with mainstream commercial applications such as a tipping system in social media, like the Dogetipbot [5] used in Twitch and Reddit platforms.

Dogecoin [6] was created by Billy Markus and Jackson Palmer to be a peer-to-peer digital currency that could reach wider audiences than Bitcoin. Being one of the cryptocurrencies that has recently grown significantly, it has reached considerably high rates in the first months of 2021. An advantage of Dogecoin over Bitcoin is that it takes only one minute for the transactions to be mined with Dogecoin, while Bitcoin miners take about 10 minutes, and the transaction fees to use Dogecoin are much lower than those for Bitcoin [7].

Dogecoin has also been predicted to grow with high confidence from the period of 2020 through 2021, due to the massive support from the investors that include celebrities and businessmen. In Figure 1, the wide variance of the Dogecoin price (DOGE) over the last one month can be observed. After its official launch on December 6, 2013, it had reached a market capitalization of 86 Billion USD, as of May 08, 2021 [8]. The market capitalization value [9] is calculated by multiplying the current supply of DOGE with current circulating supply. The circulating limit for this currency is set at 127 billion, of which 113 billion coins have already been mined [10], that is, the amount of DOGE that is liquid and in circulation.



Market cap ₹6.7T

Volume (24 hours) ₹2.0T

Circulating supply 129.5B DOGE

Fig. 1. Variation of Dogecoin Prices over the period of April 2020 - May 2021 [11]

Attention mechanism [12] has improved the quality of machine translation systems through alleviating the vanishing gradient problem. This is achieved by providing a direct path to the input vectors and has also benefited the neural model to focus on the relevant parts of the input sequence as needed.

Transformer model [13] is a neural architecture that makes use of self-attention to extract the features that are essential to the time series. It comprises of an encoder-decoder architecture [14]. The encoder transforms the input sequence to a fixed length intermediate sequence which then be decoded to form the output sequence. Attention has been employed in three places [15] in a Transformers model: Self-attention is used in the encoder and the decoder unit of the transformer, and a separate encoder-decoder-attention in the decoder unit.

Self-attention [16] can be of two types : single-head and multi-head. Single-head mechanism helps determine attention weights that help model the changes in the time series. In multi-head attention models, the attention module repeats its computations multiple times in parallel. The independent attention outputs are then concatenated and transformed linearly into the expected dimension by the multi-head attention module. Intuitively, multiple attention heads allows for mapping longer-term dependencies and shorter-term dependencies of the sequence differently.

In this paper, the hour-by-hour closing prices of Dogecoin is predicted by modelling the time series data using a multi-head self-attention transformer model. The variations in the data are captured using vector embeddings and the multi-head attention mechanism helps to map variations over multiple steps at a time. The performance of the model is evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE) and Predictive R-Square. The accuracy

978-1-6654-4112-4/21/$31.00 ©2021 IEEE

of the model is also compared against state of the art cryptocurrency prediction models. Through this work, we aim to reduce the human factors in complex systems by simplifying the association between the market price forecasting and human dependencies. The cohesion between big data system analysis and manual validation is proposed to be automated to reduce the errors in processing and enhance the system's performance.

The rest of the paper is organized as follows. Section II gives the summary of some of the best works dealing with the price prediction of cryptocurrencies in the world. The design of the price prediction model for Dogecoin using multi-attention based transformer model is given in Section III, followed by the implementation details in Section IV. The results and their analysis obtained from the implemented system is present in Section V. The summarization of the project and the proposed future works is given in Section VI and Section VII respectively.

II. Related Works

In this section, some of the best works in the price forecasting of cryptocurrencies and multi-head attention modelling models are summarized.

Traditional Deep Neural modelling architectures have performed significantly better with attention mechanisms in the aspect of predicting stock market prices. LSTM-attention models have been implemented by Zhang and Zhang [17] using k-fold cross validation on the stock prices dataset. Information screening by filtering out the high-value information from the large dataset, helps give attention to the features at different locations to capture the influence of the time period. The low accuracy and stability issues observed in the traditional deep neural models were thus overcome by LSTM-attention and Transformer models by the attention mechanism.

Transformer with a causal convolutional network for feature extraction in combination with masked self-attention was employed by Wallbridge [18] for the prediction of price movements from limit order books. The implemented model updated its learned features based on relevant contextual information from the data with a causal and relational inductive bias. With hyperparameter fine tuning, the algorithm is efficient at performing on very large datasets with specified window size.

Stock movement prediction was performed with hierarchical multi-scale Gaussian system Transformers by Ding et al [19]. An orthogonal regularization module was used here to inhibit the learning of redundant heads in the multi-head self-attention model by enhancing the diversity between each head. Compared to the existing models, such models have the ability to mine extremely long-term dependencies unlike the LSTM models.

Dilated attention augmented causal convolutional networks for stock movement prediction was implemented by Daiya et al [20] by integrating heterogeneous data sources. Event-knowledge representations from the news data by capturing direct and inverse relationships among the vent tuples to infer inter-day relationships among the data features. The attention augmented neural tensor network that is used here to extract features from the embedding generated by inverse embedding combines the embedding features learnt from heterogeneous data sources. This novel integrative approach to effectively blend views from the news and the stock price series data is done to capture the temporal dynamics of the financial integrators for price prediction.

In summary, the transformer model used multiple attention modules in parallel along with the Time2vec concept for identifying the periodic and non-periodic elements of the time-series data. The periodic elements are used for the identification of the variation of the time-series data non-periodic elements are used for predicting the probability of the output. This concept creates a temporal vector to pass to the constructed transformer model, without which the transformer will not be able to identify the relation between each time sequence. The transformer uses a multi-head attention mechanism to map multiple steps of the time series at the same time thereby capturing relationships within the long historical time series. The trained model is evaluated with multiple evaluation metrics including accuracy, MAPE and RMSE values.

III. System Design

In this section, the design of the transformer model for the prediction of Dogecoin prices is discussed.

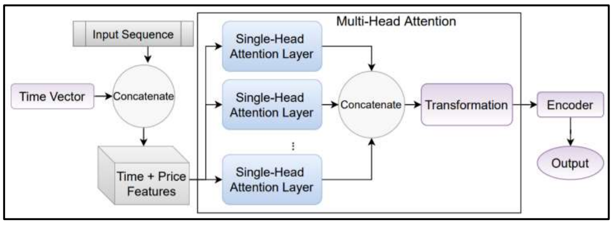


Fig. 2. Architecture of the Multi-Head Attention based Transformer Model for Price Prediction

Figure 2 shows the overall architecture of the transformer model. The input sequence is converted into a vector with time embeddings before being passed to a multi-head attention model. The features extracted are transformed and fed to an encoder model whose output is modelled to predict future prices.

A. Dataset Description

Historical trading data for Dogecoin is extracted from [21] for the period from 05 July 2019 to 28 April 2021. Open, High, Low and Close (OHLC) prices as well as the volume of Dogecoin traded is retrieved and it consists of minute-by-minute data. The data is restructured by applying hour by hour sampling to reduce number of input instances and improve the training speed of the model. In this paper, the transformer model is used to predict the closing price of Dogecoin at each hour.

B. Sequence Transformation

Sequence modelling helps identify the temporal association present between the instances in a time series. By creating a sliding window [22] of prices, input and the corresponding output sequences are generated. For each input sequence of n prices, the $(n+1)^{th}$ price is considered to be the output.

C. Time Embedding

In order to map the temporal relationships between each input sequence, transformers require embedded time vectors [23]. Without time embeddings, the transformer model would not be able to order each input sequence temporally.

Time2Vec [24] is used to generate the time embedded vectors for the input sequences.

The Time2Vec representation of the input sequence will include both periodic and non-periodic patterns [25] and the representation should be invariant with respect to time rescaling. The vector representation is given by (1) for a time series τ.

$$t2v(\tau)[i] = \begin{cases} \omega_i\tau + \varphi_i, & i = 0 \\ F(\omega_i\tau + \varphi_i), & 1 \le i \le k \end{cases} \tag{1}$$

where the periodic wave lies between $[1, k]$ and the non-periodic feature is a linear function with slope ω and intercept φ. The periodic feature comprises of a linear function which is wrapped within a function F. According to [24] the sine function performs best in terms of accuracy and hence that is chosen for the periodic feature.

The sequential input price is passed through the $t2v$ function and the periodic as well as non-periodic features are calculated. The features are concatenated with the input price sequence and are fed as a matrix to the transformer model.

D. Transformer

The transformer model forms the encoder part of the system which maps the input sequences to a fixed length sequence by applying a self-attention mechanism. Transformers help retain connection to past input sequences which are selectively determined through self-attention mechanism in order to avoid exploding the number of inputs connected. Encoder-decoder architecture is used by the transformer. The encoding mechanism involves encoding the current input based on the relevance of the current input with respect to the past inputs. This is done by passing the input through the attention weights which output sequences of variable length that are converted to fixed length by using convolutional filters.

E. Single-head Self-Attention Mechanism

Each input sequence is considered to be a token and are associated with three vectors [26] : Query (Q), Key (K) and Value (V) vectors. Query is calculated on the current token and that output is compared with the Key vector of the past tokens. The Value vector is the vector representation of the current token and it helps to determine the encoding of the token. During training of the transformer model all three vectors are optimized. The self-attention score acts as the relevance score between current token and other previous tokens. Price sequences which impact the forecasts will have a higher self-attention score compared to the other sequences which can be calculated using (2).

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where d_k is the dimension of the Key vector.

F. Multi-Head Attention Mechanism

In the Multi-Head Attention Mechanism, the attention weights of n single-head attention mechanisms are calculated, concatenated and transformed non-linearly. As there are multiple heads at the same time, the model learns multiple steps at the same time.

IV. System Implementation

The implementation details of the system are given in this section. The transformer model is set up on Google Colab with Intel(R) Xeon(R) CPU @ 2.20GHz Processor, 13 GB RAM and 12GB NVIDIA Tesla K80 graphics processor.

A. Dataset Train-Test Split

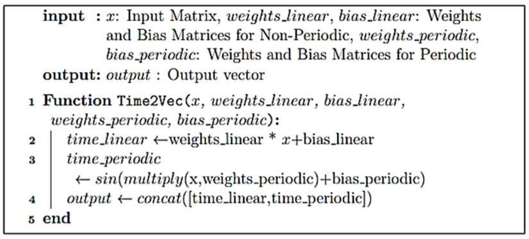
The hour-by-hour dataset collected is split into a train-test split of 80:20 as seen in Table I. A sliding window of size 50 instances is chosen to generate the input sequences and the corresponding output prediction was of sequence length 1. The data is normalized with Keras' Min-Max Scaler [27] to ensure that the data is the range of [0,1] and that the model converges faster.

TABLE I. Dataset Description

Prediction Interval	**Dataset Size**		**Sliding Window Size**
	Training Set Size	***Testing Set Size***	
Hour-by-hour	3395	849	50

B. Time2Vec Implementation

Algorithm 1: Time2Vec Embedding

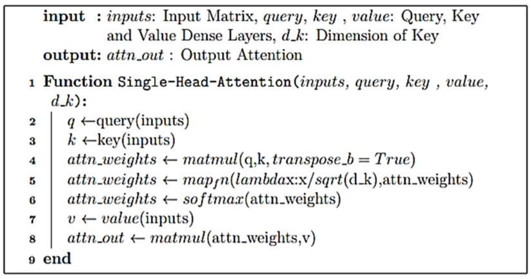


```
input  : x: Input Matrix, weights_linear, bias_linear: Weights
         and Bias Matrices for Non-Periodic, weights_periodic,
         bias_periodic: Weights and Bias Matrices for Periodic
output: output : Output vector
Function Time2Vec(x, weights_linear, bias_linear,
 weights_periodic, bias_periodic):
    time_linear ←weights_linear * x+bias_linear
    time_periodic
     ← sin(multiply(x,weights_periodic)+bias_periodic)
    output ← concat([time_linear,time_periodic])
end
```

The Time2Vec embedding matrix for the input price sequences is calculated as seen in Algorithm 1. The periodic and non-periodic vectors are calculated and are concatenated thereby providing a time dependent sequence of inputs to the transformer.

C. Calculation of Single-Head Self-Attention Weights

Algorithm 2: Single-Head Attention Mechanism



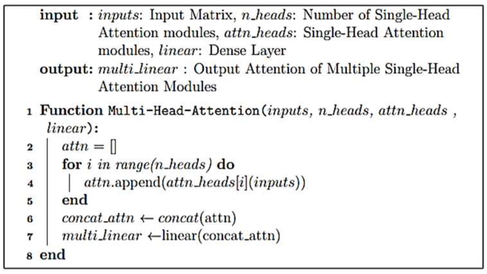
```
input  : inputs: Input Matrix, query, key , value: Query, Key
         and Value Dense Layers, d_k: Dimension of Key
output: attn_out : Output Attention
Function Single-Head-Attention(inputs, query, key , value,
 d_k):
    q ←query(inputs)
    k ←key(inputs)
    attn_weights ← matmul(q,k, $transpose\_b = True$)
    attn_weights ← $map_fn$(lambdax:x/sqrt(d_k),attn_weights)
    attn_weights ← softmax(attn_weights)
    v ← value(inputs)
    attn_out ← matmul(attn_weights,v)
end
```

The Single-Head Self-Attention weights is calculated as seen in Algorithm 2. The Query, Key, Value matrixes are generated by passing the inputs through the three corresponding Dense layers. The outputs are used to calculate the attention weights using equation (2). The Query, Key and Value Dense layers are optimized during training to reduce the loss in prediction.

D. Calculation of Multi-Head Attention Weights

The input sequence is passed through each of the n single-head attention modules, as seen in Algorithm 3, and then the outputs are concatenated together. The Dense layer in the Multi-Head Attention module performs a linear transformation of the concatenated inputs and is optimized to reduce the loss of the transformer encoder. By using multiple single-head attention models, the weights are shared between each individual model and each model tries to learn the data with multiple interpretations. There is a two level optimization present where the first level is each single-head attention model and the second level models the output from each individual model. This results in the multi-headed model having reduced loss and being a more optimized learning model.

Algorithm 3: Multi-Head Attention Mechanism



```
input  : inputs: Input Matrix, n_heads: Number of Single-Head
         Attention modules, attn_heads: Single-Head Attention
         modules, linear: Dense Layer
output: multi_linear : Output Attention of Multiple Single-Head
         Attention Modules
Function Multi-Head-Attention(inputs, n_heads, attn_heads ,
  linear):
    attn = []
    for i in range(n_heads) do
        attn.append(attn_heads[i](inputs))
    end
    concat_attn ← concat(attn)
    multi_linear ←linear(concat_attn)
end
```

E. *Transformer Encoder Architecture*

Table II specifies the training parameters for the multi-head attention based transformer model. There are 3 multi-head attention modules each with 12 heads. By increasing the number of heads, the ability to identify variations in time series improves. With 3 such multi-head modules, it is ensured that the dependencies of long historical data can be modelled efficiently.

TABLE II. TRAINING PARAMETERS OF THE TRANSFORMER

Parameter	***Hour-by-hour***
Number of Heads	12
Number of Encoders	3
Loss	MSE
Optimizer	Adam
Epochs	2000
Batch Size	256
Key Vector Size	256
Value Vector Size	256
Activation Function	ReLu
Output Activation	Linear

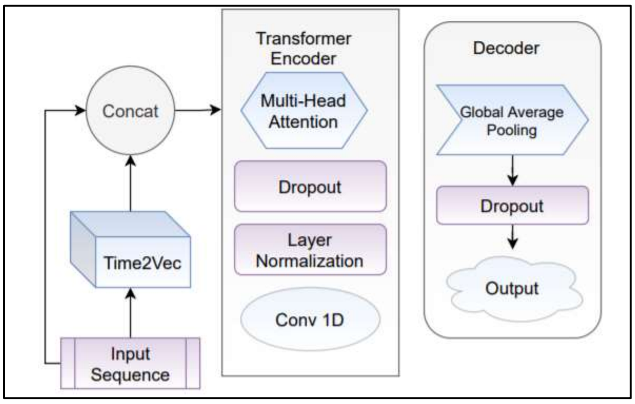


Fig. 3. Implementation of Transformer Model

Figure 3 shows the implemented transformer encoder and decoder model. The encoder unit consists of the multi-head attention model followed by Dropout, Layer Normalization and Conv1D. The convolution layer has 256 filters with a kernel size 1. They are activated by ReLu [28] activation function. The input to the encoder is the input sequence embedded using the Time2Vec vectorization. The fixed length output from the encoder is then passed to the decoder which comprises Global Average Pooling, Dropout and Dense output. The output layer is just 1 node corresponding to the output window length and is activated by using the Linear activation function [28] and it is chosen because regression problems involve predicting a numerical value instead of a probability. The linear activation function forwards the activation value that is the weighted sum of the input sequence.

V. RESULTS AND ANALYSIS

The results obtained from the testing of the trained multi-attention based transformer model and their analyses are summarized below.

A. *Training plot of the Multi-head attention based transformer model*

The constructed transformer model is trained for 2000 epochs over the hour-by-hour Dogecoin price variation data. The model is monitored with the decreasing Mean Square Error (MSE) [29] value at every 200 epochs to avoid overfitting of the data with the trained weights in the model. The training plot of the multi-head attention transformer model with the MAPE and Loss values are plotted in Figure 4. The loss value refers to the sum of errors between the actual and predicted value of the trained model. It is calculated for every epoch during the training of the deep neural model.

The Mean Absolute Percentage Error (MAPE) [29] is a measure of prediction accuracy of the constructed transformer for forecasting. The absolute difference between the actual value and the predicted value is calculated and summed for every forecasted point in time, divided by the total number of points as in (3).

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{y}\right| \qquad (3)$$

where, n is the number of times the summation iteration happens, y is the actual value and $\hat{y}$ is the predicted value.

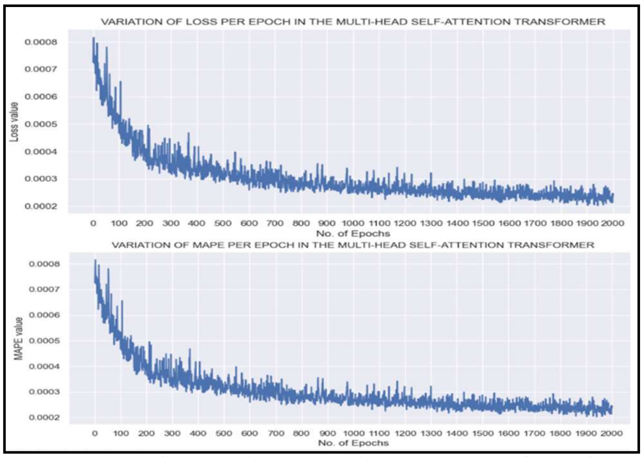


Fig. 4. Variation of Loss and MAPE in the Training Plot of Transformer model

B. *Real and prediction value comparison of the Transformer model trained with hourly data*

The actual hour-by-hour price variation of the Dogecoin is plotted along with the predicted value for comparison in the Figure 5. It is observed that the Mean Squared Error (MSE) value is at 1.36e-09 after the forecasting transformer model has completed its training. The training of the constructed deep neural model was stopped before the model overfit the data to learn the noise and other inconsistencies present in the data.

C. *Value of Accuracy, RMSE, MAE and predictive R-squared value of the Transformer model*

The trained multi-head self-attention transformer is tested with below evaluation metrics and the results are recorded in Table III.

Root Mean Square Error (RMSE) [29] value, gives the standard deviation of the residuals, which is how far the predicted data points are from the actual data points for evaluation using (4). Here, y represents the actual value, $\hat{y}$ refers to the predicted value by the model and n is the total number of observations.The RMSE values here reach a minimum value owing to the smaller quantitative values considered in the dataset.

$$RMSE = \sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad (4)$$

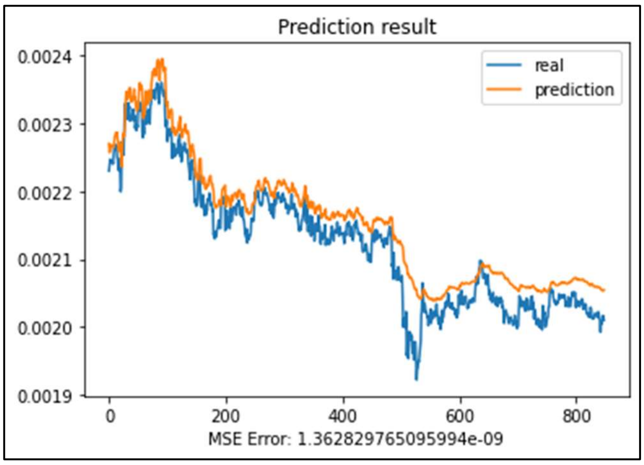


Fig. 5. Prediction result with the MSE error for hour-by-hour Dogecoin price

Mean Absolute Error (MAE) [29] is the absolute measure of errors between any paired observations expressing the same recorded phenomenon using (5). It is observed that the trained transformer model with attention mechanism is able to predict the price with a high level of accuracy and low rate of error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (5)$$

Accuracy [30] is used in a vital measure used in the evaluation of the performance of the deep neural model. Here, the value of accuracy of the constructed transformer model is calculated using (6). The accuracy of prediction indicates the percentage of deviation of the predicted values from the actual values.

$$Accuracy \triangleq \frac{1}{n}\sum_{t=1}^{n}(1 - 100 * \frac{|y[t]-\hat{y}[t]|}{y[t]}) \qquad (6)$$

TABLE III. EVALUATION METRICS OF THE MULTI-HEAD SELF-ATTENTION TRANSFORMER MODEL

Model	**Evaluation Metric**			
	RMSE	***MAE***	***Predictive R-squared value***	***Accuracy***
Multi-Head Self-Attention Transformer	25.873	3.217	86.17	98.47

Predictive R-squared [31] metric is used by systematically removing each observation from the chosen data set, estimating the regression equation and then determining the prediction performance of the trained model for the removed observation using (7), where $\bar{y}$ is the mean of all the observations.

$$R^2 = 1 - \frac{(y-\hat{y})^2}{(y-\bar{y})^2} \qquad (7)$$

D. *Comparison of the Transformer price prediction model with existing state-of-the-art methods*

The performance of the constructed transformer model is compared with some of the best existing state-of-the-art models used in the price prediction of cryptocurrency is presented in Table IV. The existing methods included the LSTM-based price prediction model along with RNN (Recurrent Neural Network) and machine learning based Linear Regression model. The use of multiple attention modules employed in parallel helps the model to recognize the short-term, dependencies long-term dependencies and other relevant features for better forecasting results. It is thus observed that the implemented multi-head self-attention transformer model has outperformed the existing methods.

TABLE IV. PERFORMANCE COMPARISON OF STATE-OF-THE-ART MODELS WITH MULTI-HEAD SELF-ATTENTION TRANSFORMER

Model	**Evaluation Metric**			
	RMSE	***MSE***	***MAE***	***Accuracy***
Our Model Multi-head Self-Attention Transformer	25.873	669.41	3.217	98.47
LSTM : Shin et al. [30]	31.75	1008.06	-	95.86
RNN : Kavitha et al. [32]	95.067	9037.74	64.389	-
LSTM : Felizardo et al. [33]	76.763	8990.969	64.854	-
GRU: Rizwan [34]	-	-	-	94.70
Linear Regression : Ali and Shatabda [35]	-	-	-	96.97

VI. CONCLUSION

The time-series forecasting of the Dogecoin price is done by using a multi-head attention transformer model that uses multiple cycles of the attention module in parallel for the identification and learning of the short term and long-term dependencies. These dependencies are learnt during the training of the deep neural model for the prediction of the Dogecoin price trend over an hour-by-hour period of time. The model is tested using evaluation metrics such as Mean Absolute Error, Accuracy of the trained model and the predictive R-squared error value to compare its performance with the existing state-of-the-art models for cryptocurrency price prediction models.

VII. FUTURE WORKS

The cryptocurrency price trends have high fluctuation rates in the global market and with interest of the investors. Thus, they exhibit a number of short-term and long-term dependencies that have to be studied by the deep neural models with their specifically relevant units. While the multi-head attention modules in Transformers has helped to identify them to a large extent, further use of deep learning networking models with higher order self-attention modules would help to analyse the multiple features in the data for self-supervised prediction of price, volume and market capitalization of bitcoins and altcoins in real-time.

REFERENCES

[1] M. R. Islam, R. M. Nor, I. F. Al-Shaikhli, and K. S. Mohammad, “Cryptocurrency vs. Fiat Currency: Architecture, Algorithm, Cashflow Ledger Technology on Emerging Economy: The Influential Facts of Cryptocurrency and Fiat Currency,” in 2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M), pp. 69–73, 2018.

[2] H. Albayati, S. K. Kim, and J. J. Rho, “Accepting financial transactions using blockchain technology and cryptocurrency: A customer perspective approach,” Technology in Society, vol. 62, 2020.

[3] A. Meynkhard, “Effect of Bitcoin Volatility on Altcoins Pricing,” Software Engineering Perspectives in Intelligent Systems, pp. 652–664, 2020.

[4] I. Young, “Dogecoin: A Brief Overview & Survey,” SSRN Electronic Journal, 2018.

[5] A. L. Massanari, “Contested Play : The Culture and Politics of Reddit Bots,” in Socialbots and Their Friends: Digital Media and the Automation of Sociality, R. W. Gehl and M. Bakardjieva, Eds, pp. 110–127, 2016.

[6] J. Biasi and S. Chakravorti, “The Future of Cryptotokens,” Disruptive Innovation in Business and Finance in the Digital World, vol. , pp. 167–187, 2019.

[7] K. Li, H. Li, H. Hou, K. Li, and Y. Chen, “Proof of Vote: A High-Performance Consensus Protocol Based on Vote Mechanism & Consortium Blockchain,” in 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp. 466–473, 2017.

[8] “Dogecoin USD (DOGE-USD) price, news, quote & history – Yahoo Finance,” finance.yahoo.com. https://finance.yahoo.com/quote/DOGE-USD/ (accessed May 08, 2021).

[9] M. Praveen Kumar and N. V. Manoj Kumara, “Market capitalization: Pre and post COVID-19 analysis,” Materials Today: Proceedings, vol. 37, no. 2, pp. 2553–2557, 2020.

[10] T. Hazlewood, “Deconstructing Dogecoin — Not All Cryptocurrencies are Equal,” Medium, 2021. https://levelup.gitconnected.com/deconstructing-dogecoin-not-all-cryptocurrencies-are-equal-ee347b29700f (accessed May 08, 2021).

[11] “Dogecoin Price Chart (DOGE) | Coinbase,” www.coinbase.com. https://www.coinbase.com/price/dogecoin.

[12] X. Yin, Y. Han, H. Sun, Z. Xu, H. Yu, and X. Duan, “A Multivariate Time Series Prediction Schema based on Multi-attention in recurrent neural network,” in 2020 IEEE Symposium on Computers and Communications (ISCC), pp. 1–7, 2020.

[13] J. Liu et al., “Transformer-Based Capsule Network For Stock Movement Prediction,” in Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2019), pp. 66–73, 2019.

[14] R. M. Farsani and E. Pazouki, “A Transformer Self Attention Model for Time Series Forecasting,” J. Electr. Comput. Eng. Innovations, vol. 9, no. 1, pp. 1–10, 2021.

[15] X. Yin, Y. Han, H. Sun, Z. Xu, H. Yu, and X. Duan, “Multi-Attention Generative Adversarial Network for Multivariate Time Series Prediction,” IEEE Access, vol. 9, pp. 57351–57363, 2021.

[16] H. Abbasimehr and R. Paki, “Improving time series forecasting using LSTM and attention models,” Journal of Ambient Intelligence and Humanized Computing, 2021.

[17] S. Zhang and H. Zhang, “Prediction of Stock Closing Prices Based on Attention Mechanism,” in 2020 16th Dahe Fortune China Forum and Chinese High-educational Management Annual Academic Conference (DFHMC), pp. 244–248, 2020.

[18] J. Wallbridge, “Transformers for Limit Order Books,” arXiv:2003.00130 [q-fin], 2020.

[19] Q. Ding, S. Wu, H. Sun, J. Guo, and J. Guo, “Hierarchical Multi-Scale Gaussian Transformer for Stock Movement Prediction,” in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, pp. 4640–4646, 2020.

[20] D. Daiya, M.-S. Wu, and C. Lin, “Stock Movement Prediction That Integrates Heterogeneous Data Sources Using Dilated Causal Convolution Networks with Attention,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8359–8363, 2020.

[21] “Crypto Archive - Welcome,” . https://www.cryptoarchive.com.au/ (accessed May 09, 2021).

[22] H. G. T. Thu, T. N. Thanh, and T. L. Quy, “A Neighborhood Deep Neural Network Model using Sliding Window for Stock Price Prediction,” in 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 69–74, 2021.

[23] W. Ji, Y. Sun, T. Chen, and X. Wang, “Two-stage Sequential Recommendation via Bidirectional Attentive Behavior Embedding and Long/Short-term Integration,” in 2020 IEEE International Conference on Knowledge Graph (ICKG), pp. 449–457, 2020.

[24] S. M. Kazemi et al., “Time2Vec: Learning a Vector Representation of Time,” arXiv:1907.05321 [cs], 2019.

[25] Y. Shen, X. Jiang, Y. Wang, X. Jin, and X. Cheng, “Dynamic Relation Extraction with A Learnable Temporal Encoding Method,” in 2020 IEEE International Conference on Knowledge Graph (ICKG), pp. 235–242, 2020.

[26] A. Vaswani et al., “Attention Is All You Need,” in 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 1–15, 2017.

[27] T. Lee, V. P. Singh, and K. H. Cho, “Tensorflow and Keras Programming for Deep Learning,” Deep Learning for Hydrometeorology and Environmental Science, vol. 99, pp. 151–162, 2021.

[28] M. A. Mercioni and S. Holban, “The Most Used Activation Functions: Classic Versus Current,” in 2020 International Conference on Development and Application Systems (DAS), pp. 141–145, 2020.

[29] P. Jay, V. Kalariya, P. Parmar, S. Tanwar, N. Kumar, and M. Alazab, “Stochastic Neural Networks for Cryptocurrency Price Prediction,” IEEE Access, vol. 8, pp. 82804–82818, 2020.

[30] M. Shin, D. Mohaisen, and J. Kim, “Bitcoin Price Forecasting via Ensemble-based LSTM Deep Learning Networks,” in 2021 International Conference on Information Networking (ICOIN), pp. 603–608, 2021.

[31] W. Marbun, Suparti, and D. A. I. Maruddani, “Modeling of composite stock price index (CSPI) using semiparametric regression truncated spline based on GUI R,” Journal of Physics: Conference Series, vol. 1524, 2020.

[32] H. Kavitha, U. K. Sinha, and S. S. Jain, “Performance Evaluation of Machine Learning Algorithms for Bitcoin Price Prediction,” in 2020 Fourth International Conference on Inventive Systems and Control (ICISC), pp. 110–114, 2020.

[33] L. Felizardo, R. Oliveira, E. Del-Moral-Hernandez and F. Cozman, "Comparative study of Bitcoin price prediction using WaveNets, Recurrent Neural Networks and other Machine Learning Methods," 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), pp. 1-6, 2019.

[34] M. Rizwan, S. Narejo and M. Javed, "Bitcoin price prediction using Deep Learning Algorithm," 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), pp. 1-7, 2019.

[35] M. Ali and S. Shatabda, “A Data Selection Methodology to Train Linear Regression Model to Predict Bitcoin Price,” in 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), pp. 330–335, 2020.