**VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**FINAL PROJECT**

**BUSINESS ANALYSIS**

Instructor:

Assoc. Prof. Nguyễn Đình Thuân

Student performance: Team 3

|  |  |
| --- | --- |
| Phạm Lê Trưởng | 20522090 |
| Lương Lý Công Thịnh | 20521960 |
| Nguyễn Ngọc Tín | 20522015 |

**Ho Chi Minh City, June 21 2023**

# TEACHER'S COMMENTS

MỤC LỤC

[TEACHER'S COMMENTS 2](#_Toc138244166)

[I. INTRODUCTION 4](#_Toc138244167)

[II. RELATED WORKS 4](#_Toc138244168)

[III. MODELING 5](#_Toc138244169)

[1. LSTM 5](#_Toc138244170)

[2. CNN-LSTM 7](#_Toc138244171)

[3. Gated Recurrent Unit 9](#_Toc138244172)

[4. Bayesian Neural Network 11](#_Toc138244173)

[5. Hidden Markov Model 12](#_Toc138244174)

[6. Linear Regression 13](#_Toc138244175)

[7. ARIMA 14](#_Toc138244176)

[8. Vector Autoregression 15](#_Toc138244177)

[9. Singular Spectrum Analysis 15](#_Toc138244178)

[IV. METHOD 16](#_Toc138244179)

[V. EXPERIMENT 18](#_Toc138244180)

[1. First gold price dataset 18](#_Toc138244181)

[2. Second gold price dataset 23](#_Toc138244182)

[3. Third gold price dataset 29](#_Toc138244183)

[VI. CONCLUSION 34](#_Toc138244184)

[VII. APPLICATION DEPLOYMENT 35](#_Toc138244185)

[VIII. GROUP WORK DISTRIBUTION 36](#_Toc138244186)

[REFERENCE 39](#_Toc138244187)

# INTRODUCTION

Currently, with the world economic situation in general and Vietnam's economy in particular is complicated and difficult to understand and gold price is one of those fluctuations. According to Tuoi Tre newspaper [1], the increase in the world gold price has little impact on the domestic gold price because many sellers sell when the price increases a little, so the price of gold does not increase sharply compared to the world. In addition, according to VTV news [2], the gold price recovered thanks to the weakening USD, leading many people to increase their investment demand in precious metals. It can be seen that the gold market is volatile as Tien Phong newspaper [3] mentioned that when the price of gold increases, buyers still lose because the buying price is higher than the selling price. From the above results, in this project, our team has chosen to predict gold price using machine learning and deep learning models in the hope of helping Vietnam's economy.

# RELATED WORKS

Gold Price Forecasting Using ARIMA Model [4], Research by Banhi Guha and Gautam Bandyopadhyay. In this paper, they tested six sets of parameters p,d,q for the ARIMA model and the best results were obtained with ARIMA(1,1,1) satisfying the statistical criteria.

Forecasting Gold Price in Rupiah using Multivariate Analysis with LSTM and GRU Neural Networks [5], Research by Sebastianus Bara Primananda and Sani Muhamad Isa. In this paper, the best model to use in gold price forecasting problems over a period of less than three years is GRU. On the other hand, for periods over three years, the LSTM is displayed with greater accuracy. From that information, this study indicates that hyperparameter tuning is more effective in optimizing the LSTM Model than the GRU Model for this study the gold price prediction problem.

Modeling Gold Price via Artificial Neural Network [6], Hossein Mombeini and Abdolreza Yazdani-Chamzini. Research results show that ANN method is a powerful tool for modeling gold price and can give better forecasting performance than ARIMA method with R square of 0.965.

Gold Price Forecasting Using LSTM, Bi-LSTM and GRU [7], Research by Mustafa Yurtsever on comparing the performance of three multivariate models (LSTM, Bi-LSTM and GRU) to predict gold price using MAE, RMSE and MAPE measures. The results show that LSTM works best with batch size parameters of 128, number of epochs of 1000, resulting in MAPE = 3.18, RMSE = 61,728 and MAE = 48.85

Forecasting Gold Prices Using Temporal Convolutional Networks [8], Research by Justin Fajou and Andrew McCarren, this research undertook a comparative analysis between Temporal Convolutional Networks (TCNs), the current state of the art machine learning approaches and a traditional time series model in gold price prediction. The results demonstrated that the TCNs consistently outperformed the other approaches chosen in this study with result RMSE = 15.26, MAE = 10.05, R2 = 0.9954 and reduced the error by more 27% in comparison with the best preforming non-TCN approach.

Prediction of gold price with ARIMA and SVM [9], Research by D Makala and Z Li, In this research paper, we found out how to predict gold price using Arima and SVM model. The results of this study show that SVM(Poly) with MAPE = 2.49, RMSE = 0.0275 and R2 = 0.9978 is said to perform much better than other SVM(RBF) and Arima models.

# MODELING

## LSTM

LSTM (Long Short-Term Memory) model is a specialized neural network architecture widely used in time series processing. This model was introduced by Hochreiter and Schmidhuber in 1997 and has become one of the most important models in the field of deep learning for time series.

The LSTM algorithm was developed to address the issue of long-term information loss in traditional recurrent neural networks (RNNs). In RNNs, information can only propagate through a limited number of neurons, and it tends to vanish as the length of the sequence increases. LSTM tackles this problem by utilizing a memory cell and gates to regulate the flow of information during the processing of time series.

A diagram of a flowchart

Description automatically generated with medium confidence

*Figure 1: Architecture of the LSTM model*

The reason for using the LSTM model in time series processing is its ability to handle long-term dependencies. Thanks to the memory cell and gates, LSTM can retain important information from the past during training, enabling the model to make better predictions on long and complex sequences, the specific formula is as follows[10]:

Input gate (i): )

Forget gate (f): )

Output gate (o): )

Memory cell (C): )

Hidden state (h):

Where:

is the input at time t.

is the hidden state from the previous layer.

, , are the values of the gates at time t.

is the memory cell state at time t.

is the hidden state at time t.

, , , are weight matrices.

, , , are bias vectors.

The formulas describe how the gates and states of an LSTM are computed based on the current input and the previous state. This process allows the LSTM to process and store crucial information from the past and influence the prediction outcomes.

## CNN-LSTM

The CNN-LSTM model is a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers, used for time series problems.

The CNN-LSTM model was developed by leveraging the strengths of both CNN and LSTM. CNN was primarily developed for spatial data processing, such as images, while LSTM was designed for sequence data processing, like time series. When applied to time series problems, the CNN-LSTM model combines the spatial feature extraction capability of CNN with the sequence understanding and prediction capability of LSTM.

The CNN-LSTM model operates by using CNN layers to extract features from the input data and then passing them through LSTM layers to understand and predict the time series. The main steps involved in its operation are as follows:

• Firstly, the time series data is divided into fixed-size windows.

• Each data window is passed through the CNN layers to extract spatial features.

• The output of the CNN layers is then fed into the LSTM layers to understand and predict the time series.

• Finally, the output of the LSTM layers is used to make predictions for future time series values.

The CNN-LSTM model is commonly used in time series problems due to the following benefits:

• Spatial feature extraction: The CNN layers in the model help extract spatial features from the time series data. This is useful for processing structured spatial data like images or sound.

• Sequence understanding and prediction: The LSTM layers in the model have the ability to understand and predict time series data. LSTMs can store long-term information and learn dependencies on previous patterns in the time series.

• Combined advantages: The CNN-LSTM model combines the feature extraction capability of CNN with the sequence understanding and prediction ability of LSTM, improving prediction performance in time series problems compared to using either network alone.

A picture containing text, screenshot, diagram, line

Description automatically generated

*Figure 2: Architecture of the CNN-LSTM model*

CNN-LSTM model [11] calculation for time series forecasting:

Convolutional Layers:

• Firstly, the initial time series is divided into fixed-size windows.

• Each window is a 2D array with the vertical dimension representing the length of the time series and the horizontal dimension representing the number of features or channels.

• Apply Convolutional layers on each window to extract spatial features from the time series.

• After applying Convolutional layers, the output of each window becomes a 2D input for the LSTM layer.

Pooling Layer:

• After applying Convolutional layers, you can apply Pooling layers to reduce the spatial size of the input and increase the model's generalization.

• Pooling layers such as Max Pooling or Average Pooling perform downsampling of the input by taking the maximum or average value within a window.

Flatten Layer:

• After applying Pooling layers, the 2D input needs to be flattened into a 1D vector to be fed into the LSTM layer.

• The Flatten layer performs this operation by reshaping the input matrices into a single vector by concatenating the rows of the matrix.

LSTM Layer:

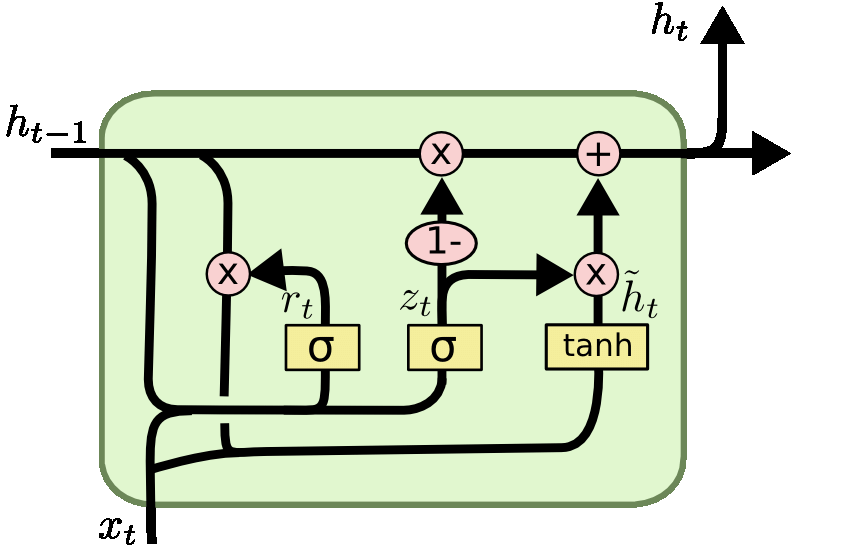
• After applying the Flatten layer, the 1D input is passed through LSTM layers to process the time series and make predictions.

• The LSTM layer has the ability to retain and learn information from the past in the time series for forecasting future values.

• The final output of the LSTM layer can be passed to fully connected layers or other layers to make the final prediction.

## Gated Recurrent Unit

The GRU model has a gate mechanism to regulate the flow of information so as to remember context in multiple time steps (Cho et al., 2014). It uses an update gate and reset gate to determine what past information can be kept or forgotten. While GRU is similar to LSTM, it combines LSTM's forget and input gates into a single update gateway. The update gate decides how much past information is passed on while the reset gate decides how much is discarded. Figure 2 shows the structure of a GRU unit. GRU outperforms LSTM in terms of training time and prediction accuracy due to its very simple structure (Jianwei et al., 2019).



*Figure 3: Architecture of the GRU model*

There are two gates in the GRU: the update and reset gates. These two gates, which decide what information is transmitted to the output, decide the information to be deleted and transferred. These operations are performed by the following equations[12]:

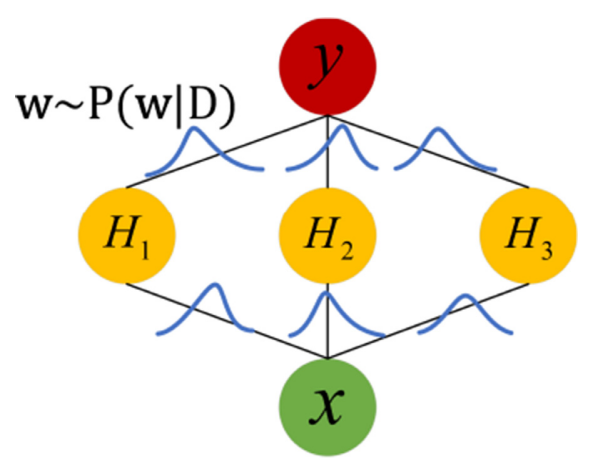
and represent the output of the current and previous states, respectively while and indicate the reset and update gates, respectively. a is the logistic sigmoid function while , and are the weight matrices.

These equations describe how the update gate, reset gate, and hidden state of a GRU are computed based on the current input and the previous state. The GRU's gating mechanisms allow it to selectively retain important information and adapt to changing input patterns.

The GRU model offers a trade-off between complexity and performance compared to the LSTM. It is computationally efficient and has fewer parameters, making it suitable for various time series processing tasks, including those involving long-term dependencies.

## Bayesian Neural Network

In traditional neural networks, each parameter has a fixed value and is determined by the classical error backpropagation. In Bayesian Neural Networks, each parameter has a posterior distribution instead of a fixed value, which is obtained by bayes backpropagation, so that the uncertainty can be introduced to the neural network prediction. Neural network based on bayes backpropagation is shown in image below, where 𝐰 is the neural network weight, and the black curves are the neural network connections. The distribution of 𝐰 is 𝑃 (𝐰|𝐷).



*Figure 4: Neural network based on bayes backpropagation*

The posterior distribution on weights 𝑃 (𝐰|𝐷) cannot be obtained directly, so researchers use the distribution 𝑞(𝐰|𝜃) to approximate 𝑃 (𝐰|𝐷) through variational learning (Graves, 2011). The variational approximation finds the parameter 𝜃 of 𝑞(𝐰|𝜃) that minimizes the Kullback-Leibler (KL) divergence with the true bayesian posterior on weights 𝑃 (𝐰|𝐷):

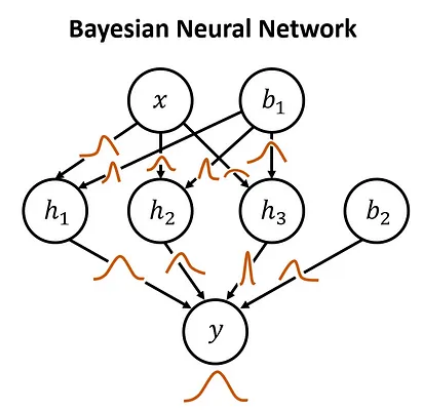
𝜃∗ =

=

=

The loss of BNN is composed of a prior-dependent part and a data dependent part. The prior distribution of parameters is usually an independent gaussian prior with mean and variance :

𝑃(𝐰) = ( |, )



*Figure 5: Architecture of the BNN model*

## Hidden Markov Model

The Hidden Markov model (HMM) is a statistical model that was first proposed by Baum L.E. (Baum and Petrie, 1966) and uses a Markov process that contains hidden and unknown parameters.

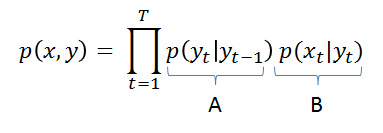
In this model, the observed parameters are used to identify the hidden parameters. These parameters are then used for further analysis. The HMM is a type of Markov chain. Its state cannot be directly observed but can be identified by observing the vector series.

Since the 1980s, HMM has been successfully used for speech recognition, character recognition, and mobile communication techniques. It has also been rapidly adopted in such fields as bioinformatics and fault diagnosis.

The basic principle of HMM is that the observed events have no one-to-one correspondence with states but are linked to states through the probability distribution. It is a doubly stochastic process, which includes a Markov chain as the basic stochastic process, and describes state transitions and stochastic processes that describe the statistical correspondence between the states and observed values. From the perspective of observers, only the observed value can be viewed, while the states cannot. A stochastic process is used to identify the existence of states and their characteristics. Thus, it is called a “hidden” Markov model.

Statistical methods are used to build state changes in HMM to understand the most possible trends in the surveillance data. HMM can automatically and flexibly adjust the trends, seasonal, covariant, and distributional elements. HMM has been used in many studies on time series surveillance data.

The HMM model has probability of any sequence of observations occurring when following a given sequence of states can be stated as follows[14]:



in which the probabilities p( |) can be read as the probability of being currently in state given we just were in the state at the previous instant t-1, and the probability p(|) can be understood as the probability of observing at instant t given we are currently in the state

## Linear Regression

Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques.[15]

The formula for a simple linear regression is:[16]

Where:

* dependent variable
* is the intercept term.
* is the regression coefficient
* is the independent variable
* is the error ter

## ARIMA

ARIMA models (Autoregressive Integrated Moving Average) were first introduced by George Box and Gwilym Jenkins in the early 1970s. They have since become a fundamental tool in time series analysis and forecasting. ARIMA model is commonly denoted as (p, d, q):

* Auto-Regressive (AR): using a linear combination of past values of the variable. An autoregressive model of order p can be written[16]

Where is current value, are model parameters, is random error.

* Integrated (I): refers to the differentiation of the time series data.
* Moving Average (MA): uses past forecast errors in a regression-like model. “q” is the number of previous error values to consider for the forecast[16]

Where is current value, is random error and are coefficients, is constants.

## Vector Autoregression

Vector Autoregressive models were introduced by Sims in 1980, revolutionized time series analysis in economics.. They are widely used for forecasting, analyzing relationships between variables, and have applications in finance and meteorology.

VAR is a forecasting algorithm that can be used when two or more time series influence each other, i.e. the relationship between the time series involved is bi-directional.[17]

The VAR(p) model of order p can be represented in the following formula[18]:

Where:

* is a vector containing the values of the variables at time t.
* is the intercept.
* , , ..., are matrices of parameters.
* is the vector of random errors.

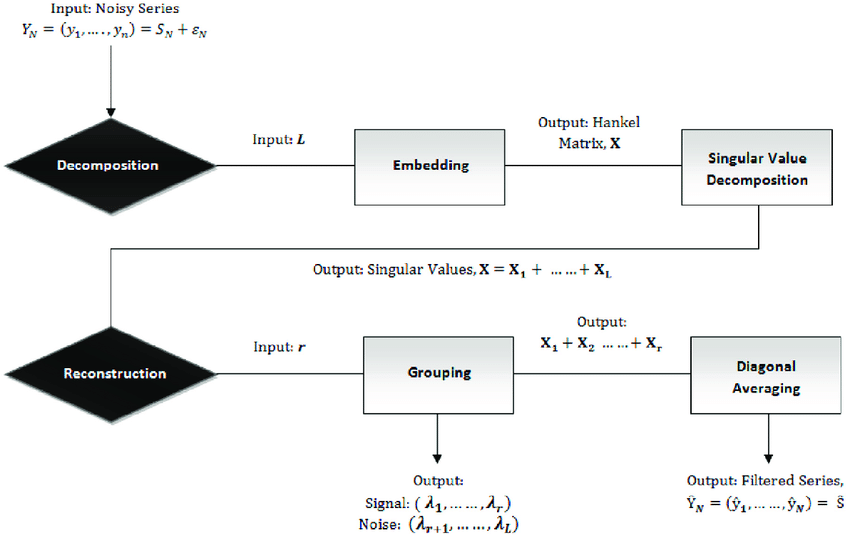
## Singular Spectrum Analysis

SSA (Singular Spectrum Analysis) is a non-parametric time series analysis method used to analyze and predict the components in a time series.After the books that were published by Vautard et al. (1989) SSA became widespread in the field of climatology and the book by Elsner (1996) shortly after, combined the prevalent information from other books related to SSA aiding the scientific communities and research institutions.[19]

The basic SSA methodology as applied to a given time series consists of two important stages[20]:

(1) Decomposition: comprises embedding and singular value decomposition (SVD)

(2) Reconstruction: comprises eigentriple grouping and diagonal averaging.



*Figure 6: Flowchart of the SSA model*

The VAR model operates as follows:

* Embedding step: the observed time series is mapped into a trajectory matrix (Hankel matrix)
* Lagged covariance matrix defined as is decomposed into eigentriplets (equal to win- dow length) by using SVD.
* Choosing suitable eigenvalues and corresponding eigenvectors for trend extraction from SVD and subsequent Hankelization matrix from selected components of the SVD.
* In grouping step, apply eading eigenfunction to select components explain the long-term trends in the original timeseries.
* Final, reconstruction of a time series by averaging the diagonal elements of selected matrices in the grouping stage.
* The PCs and corresponding eigenvectors are then considered to reconstruct the time series trend by the method of diagonal averaging.

# METHOD

The problem-solving method needs to go through the process, so we will build a framework diagram to represent the research method.

A picture containing text, diagram, screenshot, plan

Description automatically generated

*Figure 7: Diagram of research steps*

They are divided into steps:

Step 1: Data preprocessing

Step 2: Split the dataset

Step 3: Build the model

Step 4: Train the model

Step 5: Evaluate the model

Model evaluation metrics[21]: RMSE, MAPE, MAE

RMSE =

MAPE =

MAE =

Where N is the number of data points, is a predicted value and is a real value

# EXPERIMENT

## First gold price dataset

1. Data description

The dataset provided by kaggle[X] consists of 2680 rows, and time data columns and multi-country gold price columns.

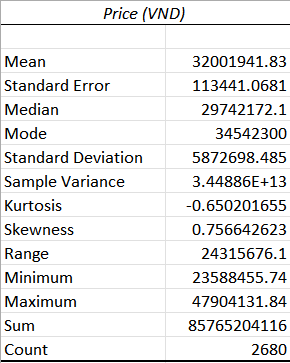
A picture containing screenshot, symmetry, pattern

Description automatically generated

*Figure 8: Original dataset*

1. Data preprocessing

* Keep the time series column and the gold price column of Viet Nam then normalize and rearrange the time series column in order
* Descriptive statistics:



*Figure 9: Descriptive statistics*

* Visualize data

A picture containing screenshot, text, plot, line

Description automatically generated

*Figure 10: Dataset visualization*

1. Split the dataset

We divide into 3 datasets: train data, test data and valid data in the ratio of 6:3:1 and 7:2:1 and 8:1:1 respectively.

1. Building model
2. LSTM

Architecture of the LSTM model

A screenshot of a computer

Description automatically generated with low confidence

*Figure 11: Summary of LSTM model*

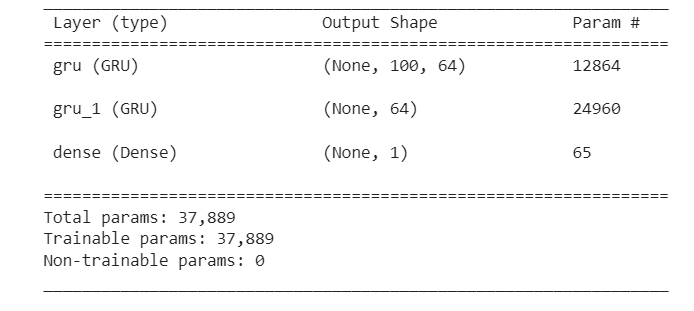
1. CNN-LSTM

A screenshot of a computer program

Description automatically generated with low confidence

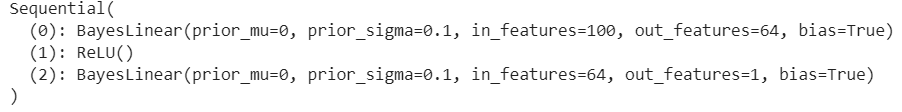
*Figure 12: Summary of CNN-LSTM model*

1. GRU



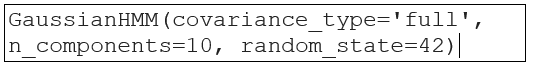
*Figure 13: Summary of GRU model*

1. BNN



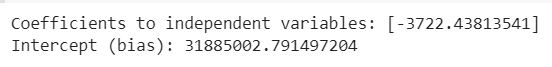
*Figure 14: Summary of BNN model*

1. HMM



*Figure 15: Summary of HMM model*

1. LN



*Figure 16: Coefficient and intercept of LN model*

1. ARIMA

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

*Figure 17: Summary of ARIMA model*

1. VAR

Ảnh có chứa văn bản, biên lai, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

*Figure 18: Summary of VAR model*

1. SSA

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

*Figure 19: Summary of embedding step*

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

*Figure 20: Summary of decomposition step*

1. Evaluate the model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train:  Test:Val | RMSE | MAPE  (%) | MAE |
| LSTM | 6:3:1 | **396747** | **0.73** | **269878** |
| 7:2:1 | 450611 | 0.78 | 312180 |
| 8:1:1 | 497241 | 0.87 | 376123 |
| CNN-LSTM | 6:3:1 | 912149 | 1.78 | 666847 |
| 7:2:1 | 892371 | 1.75 | 697204 |
| 8:1:1 | 940909 | 1.74 | 748217 |
| GRU | 6:3:1 | 516125 | 0.93 | 348108 |
| 7:2:1 | 637469 | 1.14 | 456499 |
| 8:1:1 | 548157 | 0.96 | 411844 |
| BNN | 6:3:1 | 595907 | 1.16 | 427409 |
| 7:2:1 | 813854 | 1.67 | 655251 |
| 8:1:1 | 638168 | 1.21 | 519933 |
| HMM | 6:3:1 | 14646296 | 29.55 | 11910256 |
| 7:2:1 | 6427139 | 13.27 | 5468529 |
| 8:1:1 | 4772260 | 9.35 | 3974343 |
| LN | 6:3:1 | 12445516 | 28.11 | 10602195 |
| 7:2:1 | 12712022 | 29.36 | 11612709 |
| 8:1:1 | 12780731 | 29.66 | 12581605 |
| ARIMA | 6:3:1 | 7607040 | 14.20 | 5612101 |
| 7:2:1 | 9175055 | 19.31 | 7819175 |
| 8:1:1 | 8026028 | 18.05 | 7704910 |
| VAR | 6:3:1 | 6352238 | 30.55 | 3807911 |
| 7:2:1 | 6227212 | 21.97 | 3708145 |
| 8:1:1 | 3622752 | 11.18 | 2280883 |
| SSA | 6:3:1 | 9129563 | 17.24 | 6809718 |
| 7:2:1 | 10688670 | 23.07 | 9286702 |
| 8:1:1 | 6956563 | 15.38 | 6579275 |

On the first gold price dataset, the best LSTM model results with the ratio of train, test, valid is 6:3:1 so we use this model to forecast the next 30 days

A picture containing text, screenshot, plot, line

Description automatically generated

*Figure 21: Visualize forecast results using LSTM model with first gold price dataset*

## Second gold price dataset

1. Data description

This dataset is provided by kaggle with 2290 rows and 6 columns, where one column is the time series and the other 5 columns are the values of SPX, GLD USO, SLV, EUR/USD, here we do gold price forecast should use only GLD (GOLD)

A picture containing text, screenshot, font, number

Description automatically generated

*Figure 22: Original dataset*

1. Data preprocessing

* Process and return to the form of a column of time series and a column of the gold price keep the time series column and the gold price (GLD) column then normalize and put the time series in the correct order
* Descriptive statistics:

A picture containing text, screenshot, font, number

Description automatically generated

*Figure 23: Descriptive statistics*

* Visualize data

A picture containing screenshot, plot, text, line

Description automatically generated

*Figure 24: Dataset visualization*

1. Split the dataset

We divide into 3 datasets: train data, test data and valid data in the ratio of 6:3:1 and 7:2:1 and 8:1:1 respectively.

1. Building model
2. LSTM

Architecture of the LSTM model

A screenshot of a computer

Description automatically generated with low confidence

*Figure 25: Summary of LSTM model*

1. CNN-LSTM

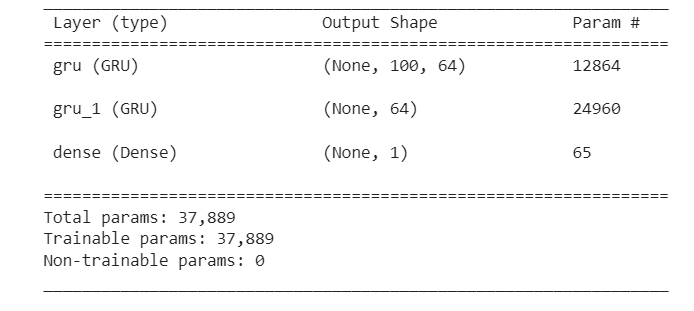
Architecture of the CNN-LSTM model:

A screenshot of a computer program

Description automatically generated with low confidence

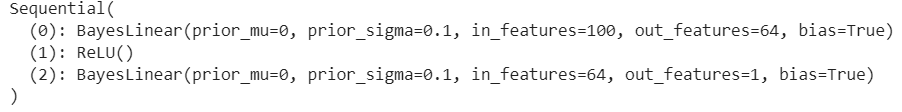
*Figure 26: Summary of CNN-LSTM model*

1. GRU



*Figure 27: Summary of GRU model*

1. BNN



*Figure 28: Summary of BNN model*

1. HMM



*Figure 29: Summary of HMM model*

1. LN



*Figure 30: Coefficient and intercept of LN model*

1. ARIMA

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Mô tả được tạo tự động

*Figure 31: Summary of ARIMA model*

1. VAR

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

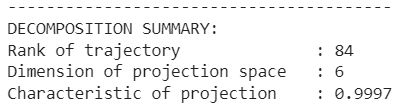
*Figure 32: Summary of VAR model*

1. SSA

Ảnh có chứa văn bản, Phông chữ, biên lai, màu trắng

Mô tả được tạo tự động

*Figure 33: Summary of embedding step*



*Figure 34: Summary of decomposition step*

1. Evaluation the model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train:Test:Val | RMSE | MAPE(%) | MAE |
| LSTM | 6:3:1 | 1.2 | 0.76 | 0.9 |
| 7:2:1 | 1.2 | 0.79 | 0.9 |
| 8:1:1 | **0.9** | **0.6** | **0.7** |
| CNN-LSTM | 6:3:1 | 1.7 | 1.16 | 1.3 |
| 7:2:1 | 1.6 | 1.05 | 1.2 |
| 8:1:1 | 1.5 | 1.04 | 1.2 |
| GRU | 6:3:1 | 1.26 | 0.08 | 0.96 |
| 7:2:1 | 1.56 | 1.06 | 1.24 |
| 8:1:1 | 1.03 | 0.9 | 1.04 |
| BNN | 6:3:1 | 2.65 | 1.8 | 2.1 |
| 7:2:1 | 2.58 | 1.76 | 2.07 |
| 8:1:1 | 2.37 | 1.77 | 2.02 |
| HMM | 6:3:1 | 31.77 | 22.39 | 25.71 |
| 7:2:1 | 31.26 | 23.5 | 27.26 |
| 8:1:1 | 14.04 | 9.6 | 11.74 |
| LN | 6:3:1 | 233 | 11.32 | 211 |
| 7:2:1 | 77 | 3.23 | 58 |
| 8:1:1 | 95 | 4.28 | 77 |
| ARIMA | 6:3:1 | 12 | 8.90 | 10 |
| 7:2:1 | 7 | 5.27 | 6 |
| 8:1:1 | 7 | 3.81 | 5 |
| VAR | 6:3:1 | 25 | 81.82 | 23 |
| 7:2:1 | 19 | 60.35 | 17 |
| 8:1:1 | 9 | 19.98 | 7 |
| SSA | 6:3:1 | 23 | 19.38 | 22 |
| 7:2:1 | 7 | 4.98 | 6 |
| 8:1:1 | 7 | 4.3 | 5 |

On the second gold price dataset, the best LSTM model results with the ratio of train, test, valid is 8:1:1 so we use this model to forecast the next 30 days

A picture containing text, plot, line, diagram

Description automatically generated

*Figure 35: Visualize forecast results using LSTM model with second gold price dataset*

## Third gold price dataset

1. Data description

The dataset provided by Yahoo Finance contains 1762 rows and is a collection of data on gold futures prices on the COMEX market in the United States, with the ticker symbol GC=F.

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Mô tả được tạo tự động

*Figure 36: Original dataset*

1. Data preprocessing

* Descriptive statistics:

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Mô tả được tạo tự động

*Figure 37: Descriptive statistics*

A picture containing screenshot, plot, line, text

Description automatically generated

*Figure 38: Dataset visualization*

1. Split the dataset

We divide into 3 datasets: train data, test data and valid data in the ratio of 6:3:1 and 7:2:1 and 8:1:1 respectively.

1. Building model
2. LSTM

Architecture of the LSTM model

A screenshot of a computer

Description automatically generated with low confidence

*Figure 39: Summary of LSTM model*

1. CNN-LSTM

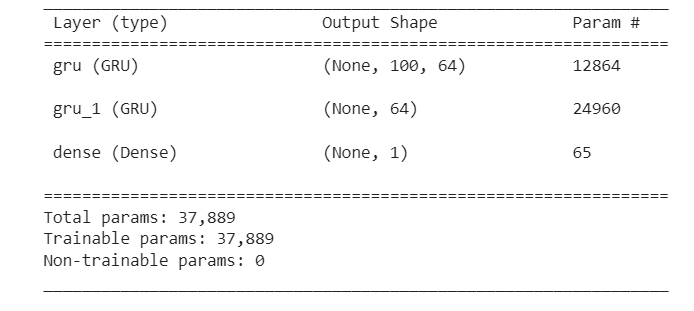
Architecture of the CNN-LSTM model:

A screenshot of a computer program

Description automatically generated with low confidence

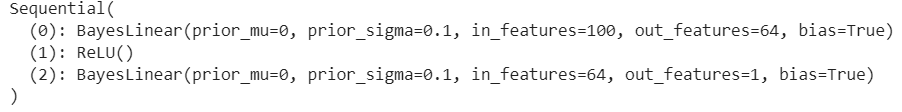
*Figure 40: Summary of CNN-LSTM model*

1. GRU



*Figure 41: Summary of GRU model*

1. BNN



*Figure 42: Summary of CNN-LSTM model*

1. HMM



*Figure 43: Summary of HMM model*

1. LN



*Figure 44: Coefficient and intercept of LN model*

1. ARIMA

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Mô tả được tạo tự động

*Figure 45: Summary of ARIMA model*

1. VAR

Ảnh có chứa văn bản, biên lai, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

*Figure 46: Summary of VAR model*

1. SSA

Ảnh có chứa văn bản, biên lai, Phông chữ, màu trắng

Mô tả được tạo tự động

*Figure 47: Summary of embedding step*

Ảnh có chứa văn bản, Phông chữ, biên lai, màu trắng

Mô tả được tạo tự động

*Figure 48: Summary of decomposition step*

1. Evaluation the model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train:Test:Val | RMSE | MAPE(%) | MAE |
| LSTM | 6:3:1 | 30.5 | 1.27 | 23.0 |
| 7:2:1 | 31.7 | 1.36 | 25.0 |
| 8:1:1 | **15.7** | **0.66** | **11.8** |
| CNN-LSTM | 6:3:1 | 62.0 | 2.93 | 52.7 |
| 7:2:1 | 35.9 | 1.51 | 27.7 |
| 8:1:1 | 29.6 | 1.34 | 23.9 |
| GRU | 6:3:1 | 20.59 | 0.83 | 15.23 |
| 7:2:1 | 20.61 | 0.88 | 16.02 |
| 8:1:1 | 25.08 | 1.12 | 20.08 |
| BNN | 6:3:1 | 72.83 | 3.45 | 65.08 |
| 7:2:1 | 33.97 | 1.53 | 27.71 |
| 8:1:1 | 71.87 | 3.49 | 65.03 |
| HMM | 6:3:1 | 424.27 | 21.75 | 392.34 |
| 7:2:1 | 148.42 | 6.7 | 121.81 |
| 8:1:1 | 218.24 | 10.21 | 182.37 |
| LN | 6:3:1 | 233 | 11.33 | 211 |
| 7:2:1 | 76 | 3.23 | 59 |
| 8:1:1 | 95 | 4.28 | 77 |
| ARIMA | 6:3:1 | 80 | 3.33 | 62 |
| 7:2:1 | 126 | 5.87 | 108 |
| 8:1:1 | 98 | 4.11 | 78 |
| VAR | 6:3:1 | 751 | 33.16 | 603 |
| 7:2:1 | 77 | 3.07 | 57 |
| 8:1:1 | 102 | 4.42 | 83 |
| SSA | 6:3:1 | 252 | 11.98 | 217 |
| 7:2:1 | 81 | 3.35 | 62 |
| 8:1:1 | 83 | 3.62 | 67 |

On the third gold price dataset, the best LSTM model results with the ratio of train, test, valid is 8:1:1 so we use this model to forecast the next 30 days

A picture containing text, plot, line, diagram

Description automatically generated

*Figure 50: Visualize forecast results using LSTM model with third gold price dataset*

# CONCLUSION

After testing nine algorithms: LSTM, CNN-LSTM, GRU, BNN, HMM, LN, ARIMA, VAR, SSA on three different datasets on gold prices, we found that the **LSTM** model is the best and gives the best results. superior results based on three measures RMSE, MAPE, MAE on all nine algorithms.

There are also some difficulties such as the training time of the model, the number of parameters used, so in the future we will improve the model for good and optimal results in terms of training time. the number of parameters used, learn new architectures such as transformer, attention, bert, .. to apply improved results

# APPLICATION DEPLOYMENT

After studying and testing the above nine models, we will use the best model is LSTM to build web applications.

Step 1: allows users to select excel or csv files to upload:

A picture containing text, screenshot, font

Description automatically generated

*Figure 51: Upload file*

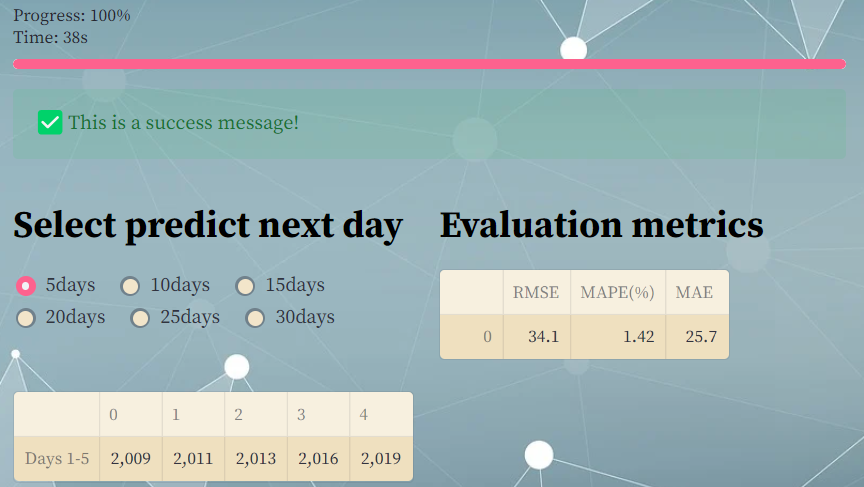
Step 2: user selects a time series column and a price series column then clicks 'Train model'

A screen shot of a computer

Description automatically generated with low confidence

*Figure 52: The two column option is the time series and the price series*

Step 3: After the training model is completed, it will show the user three evaluation measures RMSE, MAPE(%), MAE along with options to predict the next days such as 5 days, 10 days, 15 days, ... 30 days. A graph that visualizes the resulting data for the above options



*Figure 53: Price prediction results for the next 5 days and three evaluation metrics*

A picture containing text, screenshot, line, plot

Description automatically generated

A picture containing text, line, diagram, plot

Description automatically generated

*Figure 54: Visual results forecast model next 5 days*

# GROUP WORK DISTRIBUTION

|  |  |  |  |
| --- | --- | --- | --- |
| **Member**  **Work** | **Phạm Lê Trưởng (Leader)** | **Lương Lý Công Thịnh** | **Nguyễn Ngọc Tín** |
| Problem statement | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| Build a report template | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| Data collection | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| Visualizing data | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| Data Analysis | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| Data preprocessing | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| Models research | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| LSTM, CNN-LSTM model | Checkmark with solid fill |  |  |
| GRU, BNN, HMM model |  | Checkmark with solid fill |  |
| LN, ARIMA, VAR, SSA model |  |  | Checkmark with solid fill |
| Compiling codes | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| Write report | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| Build web applications | Checkmark with solid fill |  |  |
| Summarize and edit reports | Checkmark with solid fill |  |  |

# REFERENCE

[1] Vàng Thế Giới tăng mạnh, Vàng Việt Nam không đu theo (2023) TUOI TRE ONLINE. Available at: https://tuoitre.vn/vang-the-gioi-tang-manh-vang-viet-nam-khong-du-theo-20230504193607667.htm (Accessed: 21 June 2023).

[2] Baodientuvtv (2023) Đồng USD Suy Yếu, Giá Vàng Phục Hồi, BAO DIEN TU VTV. Available at: https://vtv.vn/kinh-te/dong-usd-suy-yeu-gia-vang-phuc-hoi-20230421102057128.htm (Accessed: 21 June 2023).

[3] Nga, Q. (2023) Giá vàng bật tăng, Người mua Vẫn lỗ, Báo điện tử Tiền Phong. Available at: https://tienphong.vn/gia-vang-bat-tang-nguoi-mua-van-lo-post1539111.tpo (Accessed: 21 June 2023).

[4] Bandyopadhyay, G. (2016) ‘Gold price forecasting using Arima model’, Journal of Advanced Management Science, pp. 117–121. doi:10.12720/joams.4.2.117-121.

[5] Primananda, S.B. and Isa, S.M. (2021) ‘Forecasting gold price in rupiah using multivariate analysis with LSTM and Gru Neural Networks’, Advances in Science, Technology and Engineering Systems Journal, 6(2), pp. 245–253. doi:10.25046/aj060227.

[6] Mombeini, H. and Yazdani-Chamzini, A. (2015) ‘Modeling gold price via Artificial Neural Network’, Journal of Economics, Business and Management, 3(7), pp. 699–703. doi:10.7763/joebm.2015.v3.269.

[7] YURTSEVER, M. (2021) ‘Gold price forecasting using LSTM, Bi-LSTM and gru’, European Journal of Science and Technology [Preprint]. doi:10.31590/ejosat.959405.

[8] Justin , F. and Andrew , M. (no date) Forecasting gold prices using temporal convolutional networks - CEUR-WS.org. Available at: https://ceur-ws.org/Vol-3105/paper18.pdf (Accessed: 21 June 2023).

[9] Makala, D. and Li, Z. (2021) ‘Prediction of gold price with Arima and SVM’, Journal of Physics: Conference Series, 1767(1), p. 012022. doi:10.1088/1742-6596/1767/1/012022.

[10] Christopher , O. (no date) Understanding LSTM networks, Understanding LSTM Networks -- colah’s blog. Available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (Accessed: 21 June 2023).

[11] Phyo, P.P. and Byun, Y.-C. (2021) ‘Hybrid ensemble deep learning-based approach for time series Energy prediction’, Symmetry, 13(10), p. 1942. doi:10.3390/sym13101942.

[12] YURTSEVER, M. (2021) ‘Gold price forecasting using LSTM, Bi-LSTM and gru’, European Journal of Science and Technology [Preprint]. doi:10.31590/ejosat.959405.

[13] Zhang, X., Zou, Y. and Li, S. (2022) ‘Bayesian neural network with Efficient Priors for online quality prediction’, Digital Chemical Engineering, 2, p. 100008. doi:10.1016/j.dche.2021.100008.

[14] Lan, Y. et al. (2017) ‘Development of early warning models’, Early Warning for Infectious Disease Outbreak, pp. 35–74. doi:10.1016/b978-0-12-812343-0.00003-5.

[15] Kumari, K. and Yadav, S. (2018) ‘Linear Regression Analysis Study’, Journal of the Practice of Cardiovascular Sciences, 4(1), p. 33. doi:10.4103/jpcs.jpcs\_8\_18.

[16] Hyndman, R.J. and Athanasopoulos, G. (2021) Forecasting: Principles and practice. Melbourne: OTexts. Available at: https://otexts.com/fpp2/AR.html.

[17] Gupta, A. (2021) Vector auto-regressive (VAR) models for multivariate time series forecasting, Medium. Available at: https://medium.com/geekculture/vector-auto-regressive-var-models-for-multivariate-time-series-forecasting-106bb6f74add (Accessed: 21 June 2023).

[18] Lemya , T.A. (2021) Forecasting time series using Vector Autoregressive Model [Preprint]. doi:10.22075/IJNAA.2022.5521.

[19] Wadekar, S. et al. (2022) ‘A review on singular spectrum analysis’, 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET) [Preprint]. doi:10.1109/ccet56606.2022.10080082.

[20] Aswathaiah, U. and Nandagiri, L. (2020) ‘Extraction of nonlinear trends in time series of rainfall using singular spectrum analysis’, Journal of Hydrologic Engineering, 25(12). doi:10.1061/(asce)he.1943-5584.0002017.

[21] M, P. (2023) End-to-end introduction to evaluating Regression Models, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/ (Accessed: 21 June 2023).