Joint CSCE Construction Specialty & CRC Conference 2025
Conférence conjointe spécialisée en construction de la SCGC et CRC-2025

Montreal, Quebec
July 28-31, 2025 / *28-31 juillet 2025*

AN ARTIFICIAL INTELLIGENCE FRAMEWORK TO PREDICT THE FINANCIAL PERFORMANCE OF CONTRACTORS IN EGYPT

O. Magdy[1], O. Hosny[1], I. Abotaleb[1]

[1] Department of Construction Engineering, The American University in Cairo (AUC), Cairo, Egypt

ABSTRACT: The financial success of construction companies is crucial for the growth of the industry, which is one of the main contributors to the social and economic development of any country. Contractors are usually impacted financially by external factors such as the fluctuating material prices and volatile macroeconomic conditions, making it difficult for the company's executives to make strategic decisions. The goal of this model is to predict the stock price and return on assets of different construction companies as they are the main financial performance indicators for the contractors using external and internal factors. To develop the model, financial statements of publicly listed contractors in Egypt were gathered to obtain their financial position in the previous 10 years. The input parameters including external factors such as macroeconomic indicators and material prices, and internal factors such as financial ratios were used in the model. The data was then split into training and testing and several machine learning methods were used such as Vector Auto Regression and Long Short-Term Memory to predict the key financial indicators. The models show results with reasonable mean absolute percentage error indicating that the inclusion of material prices, macroeconomic indicators and financial ratios as inputs to the model could be used to predict the stock price and return on assets. Therefore, this model could allow decision makers in construction companies to anticipate financial challenges and make strategic decisions to optimize the cash flow and improve their profitability.

1. INTRODUCTION

One of the main objectives of any construction company is to improve their financial performance by maximizing their profit and value (Chakravarthy 1986; Zeitun and Tian 2007). Different models have been recently developed to assess and predict the financial performance to evaluate the success of the company using key factors. These models usually depend on financial data that are general for all industries and do consider industry specific dynamics such as the effect of construction material prices (Singla and Samanta 2018). Therefore, traditional financial predictive models are usually inadequate for construction companies, which requires a specialized model with its own indicators tailored for the industry (Adeleye et al. 2013; Jang et al. 2019).

This study aims to address this gap by developing a predictive model for the key financial performance indicators tailored for the construction industry in Egypt. Literature review was performed to identify the dependent variables used to evaluate the financial performance, which were the return on assets (ROA) and stock price to represent the profitability and value of investment respectively in the publicly listed construction companies. The targeted literature review focuses on key studies that are relevant to the financial performance indicators of construction companies with applications in the industry. Correlation analysis was then performed between these dependent variables and the identified independent variable to remove any variable with weak correlation.

Long short-term memory (LSTM) which is a recurrent neural network and Vector Autoregression (VAR) were applied on construction companies to analyze and predict their financial performance. LSTM was selected for this study as it captures long-term dependencies and non-linear trends, while VAR excels in linear dependencies within that data. The findings revealed the ability of using these models to predict the financial performance of construction companies using the addressed factors. By having a specific model tailored for the construction industry, this paper provides an invaluable insight for enhancing the financial performance and decision making in such a complex industry.

2. LITERATURE REVIEW

The construction industry has unique challenges and a major impact on the economy which led to several studies on the financial aspects of construction companies. The literature provides different key financial indicators that are widely used to assess the financial performance of construction firms. Return on Assets (ROA) is a reliable indicator for assessing profitability as it shows how efficient the company is in utilizing their assets to make profit (Damodaran, 2012). On the other hand, the stock price provides insights for the market on the potential of future earnings of the company and its overall financial health (Fama, 1970). These indicators are relevant to construction companies as their operations are main capital-intensive and are sensitive to changes in the market.

Several independent variables that are specific for each company have been identified as influential in their financial performance. Total assets, liabilities, and working capital shows the company's liquidity and their operational efficiency (Brigham & Houston, 2021). Volume and turnover are also critical variables that reflects the trading activity of the company, and this concerns mainly investors (Lo & MacKinlay, 1999). The impact of external factors such as macroeconomic indicators were also considered in many studies.

Construction companies are highly vulnerable to economic changes, which affects their financial stability and performance according to Kim et al. (2011). This vulnerability requires the effective cash flow management and risk mitigation if the outcomes are easily predicted (Kroes & Manikas, 2014). Factors such as inflation, USD exchange rate, monthly interest rate, and unemployment rate directly impacts many contractors as they effect their purchasing power and borrowing costs (Blanchard & Johnson, 2013). Gross domestic product (GDP) and public investment in the construction industry provides insights about the economy which influences directly the construction sector (Barro, 1990). Moreover, factors that are specific to the industry such as the construction material prices are essential for the model as they directly influence the total costs and the consequently the profitability of the (Flyvbjerg, 2014).

2.1 Problem Statement

The impact of material prices along with macroeconomic indicators and company's specific financial ratios on construction companies in a volatile industry specially in Egypt is overlooked. Understanding the relationship and predicting with reasonable error between these factors would provide financial strategies for decision makers while considering the market's volatility. Previous studies have shown applications of machine learning to forecast financial aspects of companies, however, they still lack a comprehensive framework integrating the previously mentioned factor in machine learning models such as LSTM and VAR. The relation between these variables and their overall impact collectively on the financial performance indicators would provide valuable insights for decision makers in the construction industry (Chen & Hu, 2022; Vahdatmanesh & Firouzi, 2018).

3. RESEARCH METHODOLOGY

3.1 Variable Selection

Several corporate financial performance models were conducted in different studies. The dependent variables used in this model for predicting the financial performance of construction companies were chosen by conducting a thorough literature review and expert interviews to identify the financial indicators that affect financial performance. Based on the most frequently used dependant variable as shown in Table 1, stock price was selected in 4 previous studies while ROA was used in 3 of them while others were less than them, therefore these variables were ultimately chosen to ensure a holistic assessment of the financial performance. The ROA represents the profitability relative to the size of the company which concerns the internal stakeholders while the stock price represents the investment value which concerns external investors.

Table 1: Main Financial Performance Indicator from Literature

Indicator \ Paper	Mohamed et al.	Mishra et al.	Ali et al.	Kangari et al.	Fama	Song	Assaad et al.	Damodaran
Stock Price		✓			✓	✓	✓	
Net Profit	✓		✓					
Volume	✓							
Working Capital	✓							
ROA			✓	✓				✓
current ratio				✓				
liabilities to net worth				✓				
assets to revenues				✓				
revenues to working capital				✓				

As for the independent variables that affect these indicators, a list of variables was created that were often used in related studies and were then verified by interviewing 3 different financial managers working in construction companies with more than 20 years of experience in the field. The main question asked was "What are the main external and internal variables that affect the financial performance of your company?". These interviews confirmed the usage of the factors obtained from the literature and identified other factors that were overlooked such as the construction material prices. Table 2 summarize these variables that were frequently used. These factors are going to be further shortlisted according to their correlation with the dependant variables during the data preprocessing.

Table 2 Key Variable Affecting Financial Performance of Construction Companies

Category	Variable	Description	Paper
Company specific factors	Total Asset	Total amount of assets owned by an entity	Lee et al. (2017), Tan et al. (2015), Xu et al. (2016), Brigham & Houston, 2021
	Quick Assets to Current Liability	Current ratio that measures a company's ability topay short-term liability	Cheng et al. (2015), Feidakis et al. (2007),Kangari et al. (1992), Ng et al.
	Volume	The total quantity of goods or services produced or traded within a given period	Lesniak and Juszczyk (2018), (2019), Hesami and Lavasani (2014), Lo & MacKinlay, 1999
	Turnover	total value of the sale of services or goods during a certain time	Mohamed et al. (2014), Al-Matari et al. (2014), Lo & MacKinlay, 1999
	Working Capital	The difference between current assets and current liabilities, indicating short-term	Mohamed et al. (2014), Al-Matari et al. (2014), Brigham & Houston, 2021
	Inflation	The rate at which the general price level of goods and services rises	Mohamed et al. (2014) , Al-Matari et al. (2014), Chen & Rancière (2016), Blanchard & Johnson, 2013
Macroeconomic Indicators	USD Exchange Rate	The value of one EGP relative to the US dollar	Al-Matari et al. (2014), Chen & Rancière (2016), Blanchard & Johnson, 2013
	Monthly Interest Rate	The percentage charged on borrowed money or earned on investments on a monthly basis	Al-Matari et al. (2014), Chen & Rancière (2016)
	Unemployment Rate	The percentage of the labor force that is unemployed	Chen & Rancière (2016), Blanchard & Johnson, 2013
	Government Net Lending / Borrowing	The difference between government revenues and expenditures	Chen & Rancière (2016)
	Gross Domestic Product	The total monetary value of all goods and services produced	Al-Matari et al. (2014), Chen & Rancière (2016), Barro, 1990
	GDP Growth	The percentage increase or decrease in a country's GDP	Chen & Rancière (2016), Moramarco (2021), Barro, 1990
	Public Investment in Construction	Government spending on infrastructure and construction projects	Barro, 1990
Material Prices	Round Rebar Price	The market price of round reinforcing bars per ton	Flyvbjerg, 2014
	Portland Cement Price	The market price of portland cement per sack	Flyvbjerg, 2014

3.2 Data Collection

The first step was to collect the financial statements of all publicly listed companies in Egypt to extract their historical financial performance. The software used for this process is called Refinitiv which is one of the largest providers of data for the financial markets. This was used to gather company specific indicators such as the ROA, stock price, income and other financial ratios. R programming language was then used to align the general market indicators with each of the selected variables for each company to build the database. The final dataset was adjusted to timeseries format with monthly frequencies using interpolation and monthly means where necessary. Macroeconomic indicators were obtained from the Central Bank of Egypt (CBE) and the material prices were obtained from the Central Agency for Public Mobilization and Statistics (CAPMAS). These data were available for each month from the period 2013 to 2023 which was used in the model's database.

The dataset used for this research included financial data from all 16 construction companies that are publicly listed and have data available for the last 10 years. The train-test split ratio was set at 90%-10% with the first 108 months used for training and the last 12 months used for testing. This ensures that there is sufficient training data and a reasonable period for evaluating the accuracy of prediction by the model.

3.3 Data Preprocessing

The data was appropriately pre-processed before developing the model to improve the model performance. This is done by removing any missing months to ensure that the missing data does not disrupt the sequence in the time-series data. Data scaling is then carried out to transform the features to a range between 0 and 1. This helps prevent features with larger values from dominating the model's learning process as the variables used in this model have varying magnitudes and ensures that all features contribute equally.

Scaling is done after splitting the data into training and testing to prevent data leakage. The scaling is done using the following min-max scaling method where z is the scaled value of the unscaled variable x:

[1] $z = \frac{x - \min(x)}{\max(x) - \min(x)}$

Pearson correlation test was carried out to investigate the relation between the financial performance indicators and the independent variables to identify those that have the most impact on them and eliminate those with weak impact. The value of the coefficient obtained in the Pearson correlation test indicates the strength of the relation between the variable according to the classification shown in table 3 below (Xiao et al., 2016):

Table 3: Strength of Pearson Correlation (Xiao et al., 2016)

Pearson Correlation Coefficient Value	Strength of Relationship between Variables
-1 to -0.5 / +0.5 to +1	Strong
-0.5 to -0.3 / 0.3 to 0.5	Moderate
-0.3 to -0.1 / 0.1 to 0.3	Weak
-0.1 to 0 / 0 to 0.1	None

The Granger causality test is also used to check if one time series have an effect and could be used to forecast changes in another. The test is used in this study to investigate stock price and ROA casualty on the other relevant variables.

4. MODEL DEVELOPMENT

4.1 Model 1: LSTM Model

LSTM is a recurrent neural network where the neural takes the input, goes into the activation function, then gives an output that goes through another neural and loops back to the same neural. In LSTM we can control whether it reads only short or mix of both. This is useful because in real life models, variables do not always depend on the short-term memory only but depends on both. Unlike traditional recurrent neural networks, which can struggle with issues like vanishing or exploding gradients, this model is built to handle these challenges effectively, making it more reliable for learning complex temporal relationships (Jang et al., 2020). The model learns the relationships between the key variables selected and how they evolve over time to predict their future values. It leverages the power of LSTMs to handle sequential data and a multi-output structure to capture interdependencies between features. The input features for this model were independent variables as well as the dependant variable with multi-output features. Predicting all the features together ensures that the predicted target variables in the previous step is considered in the system and would provide better correlations in the model.

The LSTM unit contains four main components: a cell, an input gate, an output gate, and a forget rate, as shown in Figure 2. The cell stores the values over time intervals, while the gates control how information flows in and out of the unit. In this paper, different combinations of hyperparameters were fine-tuned using Blocked Time-Series Cross-Validation (BTCV) to identify the configuration that results in the lowest validation error. These hyperparameters included the number of cells or neurons, the activation function, batch size, and the number of epochs. BTCV was used to divide the training dataset into 5 blocks representing a continuous time period, and ensures that the training set always comes before the validation set in terms of time. The training set was all the blocks prior to the validation block while the validation set was the subsequent block in the sequence for every fold. To avoid overfitting, early stopping was used using a 10-epoch patience limit, which meant that training would end if the validation loss did not decrease for 10 epochs in a row.

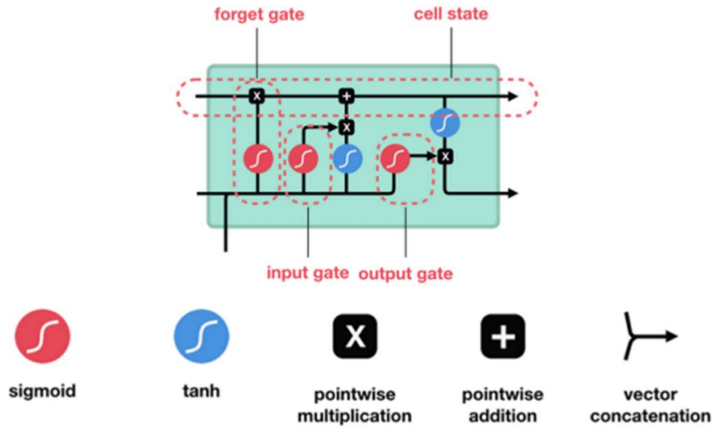

Fig. 1: Structure of an LSTM model

The model was developed using python on Google Colab. The main software package used is tensorflow which is the main library for the LSTM model that defines its architecture, compiling the model with an optimizer and loss function, and use it to make the predictions. Another essential package is the train_test_split which is used to divide the data into training and testing data sets. This allows the data to be splited into 9 years to be trained on and the last 1 year was used by the model to be trained on and assess its performance. MinMaxScaler is also used to perform the scaling of the data as mention in the data preprocessing section. Finally, numpy was used to create the array and stores the data while pandas was also used to load the dataset and perform data cleaning.

4.2 Model 2: VAR Model

VAR is a statistical method for analyzing and capturing the evolution and the interdependencies between multiple time series. In the application of this research, VAR is used to study the interaction between multiple time series simultaneously. Unlike traditional regression methods—whether linear or nonlinear—which only examine relationships between variables at the same point in time, VAR goes further by generating equations that predict the current value of a time series based on its own historical values and the historical values of other related time series, up to a specified lag order. For example, if there are two time series, y1t and y2t, and a lag order of 2, the VAR model would produce an equation for y1t that incorporates its own past values and the past values of y2t. This approach provides a more comprehensive framework for understanding and forecasting multivariate time series.

The predictions made using VAR is based on the hypothesis that the time series current state is impacted by both its own historical data and those of the other variables that are related to it. This makes VAR models more reliable than normal regression models which only is based on the relation at the same time step. However, VAR models forecast the value in a time series by using an equation including the historical data with a lag up to a predetermined lag order. Equation [3] shows an example of calculating the forecasted value with a lag of p=2 using VAR where y_{1t} is the first time series that is forecasted in this equation and y_{2t} is the other dependant time series (Lütkepohl et al. 2004).:

[1] $F(y_{1t}, y_{2t}) = a_{11} \times y_{1t-1} + a_{12} \times y_{1t-2} + a_{21} \times y_{2t-1} + a_{22} \times y_{2t-2} + c$

The VAR model was developed on R using R studio xts (Ryan and Ulrich 2018) were used to manipulate time series data. Further, the software package forecast (Hyndman et al. 2018; Hyndman and Khandakar

2007) was used to predict GDP using the fitted VAR model. Finally, the software package vars (Pfaff 2008a, b) was used to create and fit the VAR model. As for the lag order optimization, the Akaike Information Criterion (AIC) was used and the model was fitted using the least squares method.

5. MODEL EVALUATION

The models developed in this study were tested on both an unseen data of the last year and testing dataset to evaluate their performance. The mean absolute percentage error (MAPE) was used to evaluate the prediction accuracy of the model for the dependant variables chosen as shown in equation [2]. MAPE was chosen over other evaluation metrics such as RMSE or MAE as it provides the error based on percentage, which is more suitable for comparing different financial metrics across companies of varying sizes.

[2] $MAPE = \frac{100}{n} \sum_{n=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|$

The MAPE obtained for the LSTM model was better using python programming language than using R programming language and vice versa for the VAR model. The following section will show different trials for tuning the models and their performance based on the lowest MAPE.

6. RESULTS AND DISCUSSION

The Pearson correlation test showed that most of the variables selected had a strong correlation with both target variables. However, some variables such as inflation, GDP and the government gross debt had weak correlations and therefore were removed from the model. Fig. 3 shows the result obtained from the correlation analysis performed. The numbers shown range between -1 and 1 where there is a strong positive correlation if the number is closer to 1, strong negative correlation as it gets closer to -1, and no correlation if it is at 0.

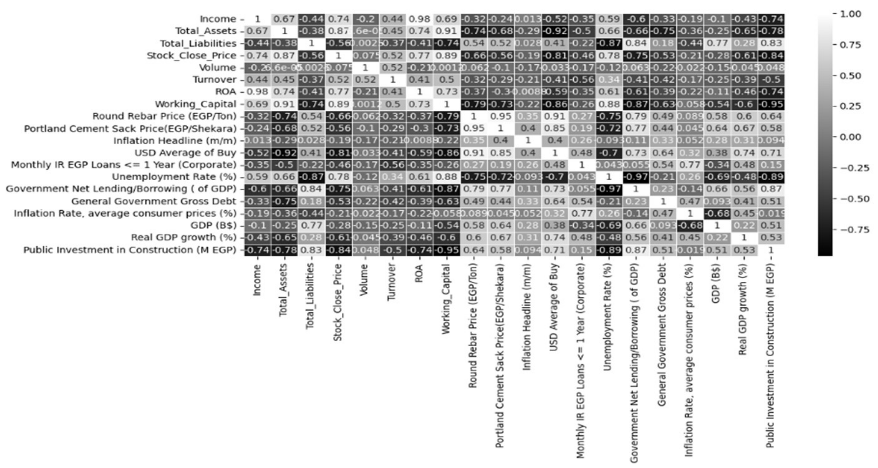

Fig. 3: Pearson corelation test performed

The remaining variables were tested for Granger causality test. Granger causality test gave p values less than 0.05 for most these variables when tested against both stock price and ROA. This means that the null hypothesis is rejected and these variables granger cause the stock price and ROA. Fig. 2 shows a sample of the Granger causality test obtained for one of the variables which is the income of the company. The p-value is significantly less than 0.05, therefore the null hypothesis that income does not Granger-cause the stock price or ROA is rejected and so it could be used in the model. Table 4 summarizes the test results for the remaining variables, and showing the variables that were removed due to a lack of causality.

```
      Granger causality H0: inc do not Granger-cause stock roa assets liab steel infl usd int unemp gdb inv

data:  VAR object Model1
F-Test = 4.3725, df1 = 66, df2 = 384, p-value < 2.2e-16
```

Fig. 2: Sample of Granger causality test

Table 4: Granger causality test results

Predictor Variable	p-value	Granger-Causes ROA/Stock Price?
Income	2.20E-16	Yes
Assets	0.0021	Yes
Liabilities	0.0053	Yes
Steel Prices	0.0012	Yes
Inflation	0.0321	Yes
USD Exchange Rate	0.0587	No
Interest Rate	0.0049	Yes
Unemployment Rate	0.118	No
GDP Growth	0.0038	Yes
Investment in Construction	0.0022	Yes

Both models were tuned with different settings and the best results for each model are reported in this section to compare the performance. The tuned models were then tested with and without including the material prices variables to investigate their impact on the results. For the LSTM model, the optimal configuration, which resulted in the lowest average MAPE across all folds, was found to be 64 neurons per layer, ReLU activation function, a batch size of 32, 50 epochs, and a learning rate of 0.001. These hyperparameter settings provided predicted data with MAPE of 8.1% for the stock price and 11.2% for the ROA when including the prices compared to having MAPE of 9.3% and 27% for the stock price and ROA respectively without including material prices. As for the VAR, the tuned model had 3 months of lag time providing predicted stock price with MAPE of 4.81% and 15.8% for the stock price and ROA respectively while including the material prices. As for the VAR model, excluding the material prices from the model provided MAPE of 8.19% and 16.1% for the stock price and ROA respectively, while adding the material prices provided predicted data with MAPE of 4.81% and 15.14% respectively.

The better performance of the VAR model to predict the stock price can be attributed to the linear movements of the stock price, which aligns with the VAR model assumptions. Even though VAR had better MAPE in predicting the stock price, the predictions do not closely follow the actual data trends for the stock price starting from Jan 2023 as shown in figures 6 and 7. This justifies that the VAR underperform during non-linear market shocks which happened during this period due to the Egyptian currency devaluation. As for the LSTM model, it can capture complex and non-linear relationships which makes it better for predicting ROA as it is influenced by a wider range of internal and external factors, so it exhibits more complex dependencies in the dataset provided.

Overall, both models showed improvement in the results when the material prices are introduced to the other variables. The VAR model had better prediction for the stock price while the LSTM model was better for the ROA. This shows that movement of the stock price tested had a linear movement making it better to use VAR. On the other hand, the ROA had more complex relationships that were captured better using the LSTM model.

Fig. 4 Actual vs Predicted on Testing Data using LSTM without Material Prices

Fig. 5 Actual vs Predicted on Testing Data using LSTM with Material Prices

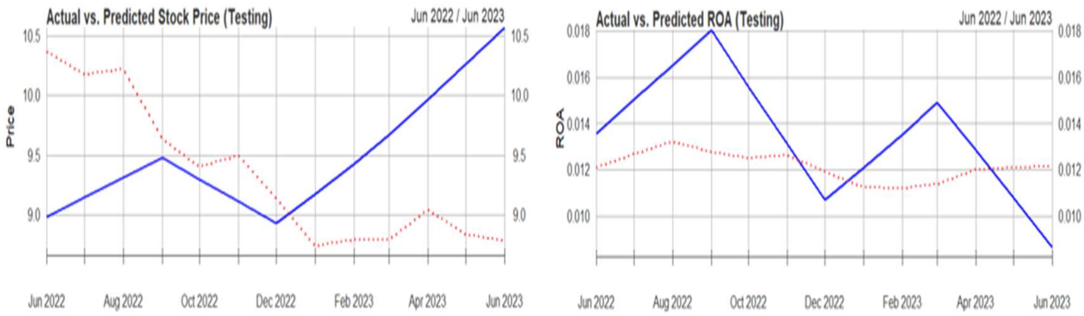

Fig. 6 Actual vs Predicted on Testing Data using VAR without Material Prices

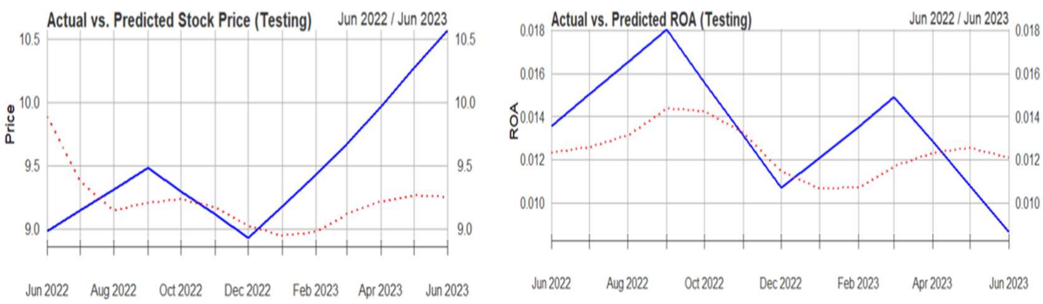

Fig. 7 Actual vs Predicted on Testing Data using VAR with Material Prices

7. CONCLUSION

This paper proposed two different AI-based predictive models to forecast the stock price and ROA which are the main financial performance indicators of construction companies in Egypt. Both models had lower MAPE after including the construction material prices as one of the dependant variables indicating that any fluctuation in these costs have an impact on the overall financial health of the company. LSTM provided better results in predicting ROA, while the VAR model provided better stock predictions for the stock price, showing that the varying complexity of the financial indicator must be considered to choose the model with better performance. This study shows the potential of AI-based models to optimize the predictions made to have more confident decision-making strategies by the stakeholders in construction companies. However, further research should be carried out with a dataset that includes the private sector firms, and a real-world validation should be conducted on different construction companies to make it more generalizable and with practical applicability. Moreover, additional forecasting models could also be tested and compared with the performance of the LSTM and VAR models as there might be additional factors that affects their performance other than the material prices.

8. REFERENCES

Al-Matari, E. M., Al-Swidi, A. K., & Fadzil, F. H. B. (2014). The Measurements of Firm Performance's Dimensions. Asian Journal of Finance & Accounting, 6(1), 24-49.

Assaad, R. & El-adawy, I., (2021). "Stock Prices of Architectural, Engineering, and Construction Firms as Leading Economic Indicator: A Computational Deep-Learning Econometrics Model to Complement the Architecture Billings Index". Journal of Architectural Engineering. Volume 27, Issue 4. https://doi.org/10.1061/(ASCE)AE.1943-5568.0000

Blanchard, O., & Johnson, D. R. (2013). Macroeconomics (6th ed.). Pearson.

Brigham, E. F., & Houston, J. F. (2021). Fundamentals of Financial Management (15th ed.). Cengage Learning.

Chakravarthy, B. S. 1986. "Measuring strategic performance." Strategic Manage. J. 7 (5): 437–458. https://doi.org/10.1002/smj.4250070505.

Chen, S., & Rancière, R. (2016). Financial Information and Macroeconomic Forecasts. IMF Working Paper, WP/16/251.

Damodaran, A. (2012). Investment Valuation: Tools and Techniques for Determining the Value of Any Asset (3rd ed.). Wiley.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417.

Kim, S., Lee, S., & Kim, J. (2011). Relationship between the financial crisis of korean construction firms and macroeconomic fluctuations. Engineering Construction & Architectural Management, 18(4), 407-422. https://doi.org/10.1108/09699981111145844

Mishra, S., & Sudarsan, P. (2019). An efficient portfolio construction model using stock price predicted by support vector regression. The North American Journal of Economics and Finance, Volume 50, 101027, 1062-9408, https://doi.org/10.1016/j.najef.2019.101027.

Mohamad, H., Ibrahim, A., & Massoud, H. (2014). Modelling the financial performance of construction companies using neural network via genetic algorithm. Canadian Journal of Civil Engineering, 41(11), 945-954. https://doi.org/10.1139/cjce-2014-0065

Jang, Y., Jeong, I., and Cho, Y. K. (2020). Business Failure Prediction of Construction Contractors Using a LSTM RNN With Accounting, Construction Market, and Macroeconomic Variables. Journal of Management in Engineering, 36(2):04019039.

Lütkepohl, H., M. Krätzig, and P. C. B. Phillips. 2004. Applied time series econometrics. Cambridge, UK: Cambridge University Press.

Siew, L., Jaaman, S., & Fai, L. (2021). Performance evaluation of construction companies using integrated entropy–fuzzy vikor model. Entropy, 23(3), 320. https://doi.org/10.3390/e23030320

Vítková, E. and Kocourkova, G. (2021). Key financial indicators by the size of the construction company - czech study. Civil and Environmental Engineering Reports, 31(2), 85-104. https://doi.org/10.2478/ceer-2021-0021

Varghese, B. and Menacere, K. (2012). The financial health of construction companies in qatar: a case study. International Journal of Engineering Research in Africa, 8, 55-72. https://doi.org/10.4028/www.scientific.net/jera.8.55

Kroes, J. and Manikas, A. (2014). Cash flow management and manufacturing firm financial performance: a longitudinal perspective. International Journal of Production Economics, 148, 37-50. https://doi.org/10.1016/j.ijpe.2013.11.008

Roozbeh Kangari, Foad Farid, Hesham M. Elgharib. (1992) Financial Performance Analysis for Construction Industry. Journal of Construction Engineering and Management. doi:10.1061/(ASCE)0733-9364(1992)118:2(349)

Song, Xinyuan, (2024) “Predicting stock price of construction companies using weighted ensemble learning”. Heliyon, Volume 10, Issue 11, e31604

Xiao, C., Ye, J., Esteves, R. M., and Rong, C. (2016) Using Spearman's correlation coefficients for exploratory data analysis on big dataset. Concurrency Computat.: Pract. Exper., 28: 3866–3878. doi: 10.1002/cpe.3745.