EasyChair Preprint

№ 7373

An Investigation of Forecasting Tadawul All Share Index (TASI) Using Machine Learning

Galal Binmakhashen, Adnan Bakather and
Ali Abdulqader Bin-Salem

January 27, 2022

An Investigation of Forecasting Tadawul All Share Index (TASI) Using Machine Learning

Galal M. BinMakhashen[+], Adnan Bakather[*]
[+]*Research Institute*
[*]*Interdisciplinary Research Center in Finance and Digital Economy*
King Fahd University of Petroleum and Minerals
Dhahran, Kingdom of Saudi Arabia
[+]binmakhashen@kfupm.edu.sa, [*]adnan.bakather@kfupm.edu.sa

Ali Abdulqader Bin-Salem
School of Computer Science and Technology
Zhoukou Normal University,
Zhoukou, Henan, 466000, China
alib.salem@yahoo.com

***Abstract*—Stock markets are one of the most complex, and dynamic environments. To make predictions about the stock prices, we may require combining several sources of market information. Another possibility is to attempt to monitor and predict the stock index prices of a target market. In this study, we investigated several machine learning algorithms to predict the Saudi stock price index by utilizing Bloomberg's most used indicators. The collected data represents 26 years of TASI index prices. Several machine learning algorithms were investigated for forecasting midterm TASI index pricing. Two Recurrent Neural Network (RNN) architectures (deeper, and shallower architectures) were created and their performances in forecasting TASI index prices are contrasted. Furthermore, several traditional machine learning methods such as Linear regression, decision trees, and random forests are also studied for index price prediction. The experiments suggested that with 26 years of TASI index transactions, simple machine learning models are generally suitable to make better midterm index price forecasting in comparison with more complex ML models.**

***Index Terms*—Stock index price prediction , Regression , Long and short-term memory network , Multivariate time series**

I. Introduction

The prediction of stock index price is a vital issue in financial market research. Index performance is subject to different investigation due to its importance to several stakeholders including markets players and policymakers [1], [2]. Market players for example, track the performance of stock index due to planing and executing investment and funding decisions. Funds managers also utilize index performance as benchmark for their investment performance. Moreover, policymakers are also interested in understanding the dynamic and movement of stock market index as important price discovery tool.

In Saudi Arabia,Tadawul All Share Index (TASI) is the most important index for the Saudi equity market as well as the exchanges in the Middle East/North Africa (MENA) region. Currently, the index traces the performance of the biggest markets in the Arab region. Traditionally, the oil price is considered the most important index determinant [3], [4]. However, the prediction of the TASI index might be still questionable in particular after the economic reform in the last five years which may change the dependence of Saudi equity markets on oil prices.

Identify applicable funding agency here. If none, delete this.

The number of studies conducted on using machine learning is still limited especially for stock index values. Besides that, due to the increased interest of investors in Islamic stocks to diversify their portfolios, there is a need to forecast the stocks performance. This work aims to collect the stock index value of Tadawul All Share Index (TASI) from the Suadi Arabia over a long period of twenty-six years and develop a robust forecasting framework for predicting the TASI index values. The index value have been analyzed using Bloomberg platform to extract several indicators. Our goal is to build a machine learning or a deep learning model to learn from these features that represent the past patterns of daily TASI index values. The learned models can be exploited in forecasting the medium-term (i.e., weekly) future index values of the TASI series with fairly high-level accuracy. We validated our experiments by by keeping the data of year-2021 for testing while building machine learning models using different time-span data (1 year, 2 years, 3 years, 4 years, 5 years, 10 years, and 25 years) in order to evaluate the predictive power of the forecasting models.

The paper is organized as follows: a background information is given in Section II. Then, related previous studies are presented in Section III. Section IV discusses the proposed framework for index prices prediction. Experimental settings and results are presented in Section V. Finally, some remarks and conclusions are presented in Section VI.

II. Background

In the financial sector, artificial intelligence (AI) including machine learning and deep learning are transformative technologies that disrupts many sectors but offer numerous potential benefits. The forecasts estimate that, by incorporating AI in its operations, the financial services sector can save a billion dollars by 2030 [5]. In general, there are three categories of forecasting methods for stock index prices as follows [6]: 1) Fundamental Analysis, 2) Technical Analysis, 3) Time Series Forecasting.

On one hand, the fundamental analysis is an investment analysis where corporate sales, earnings, profits, and all other economic factors are analyzed to draw conclusions over the investment. Hence, the fundamental analysis can be considered

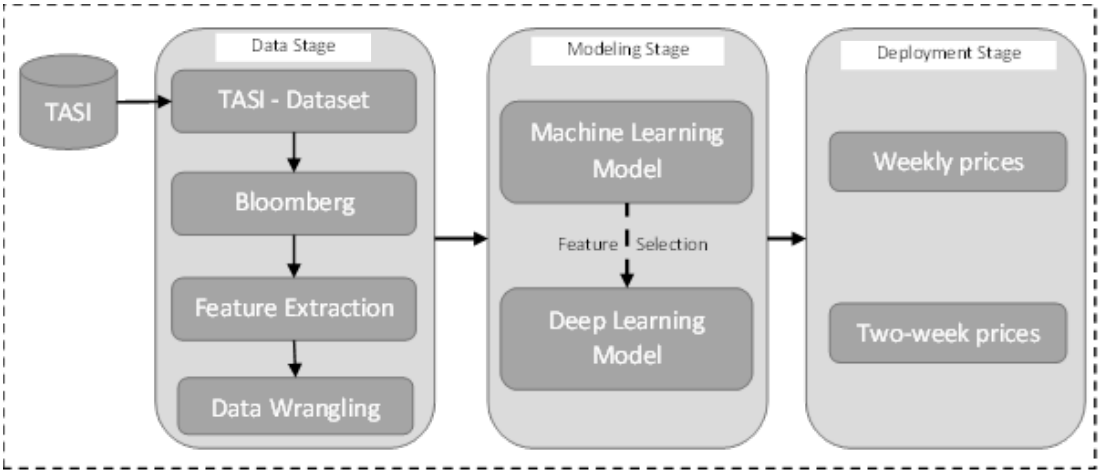


Fig. 1: Overview of the proposed framework

for long-term forecasting. On the other hand, the technical analysis uses the historical stock index prices.

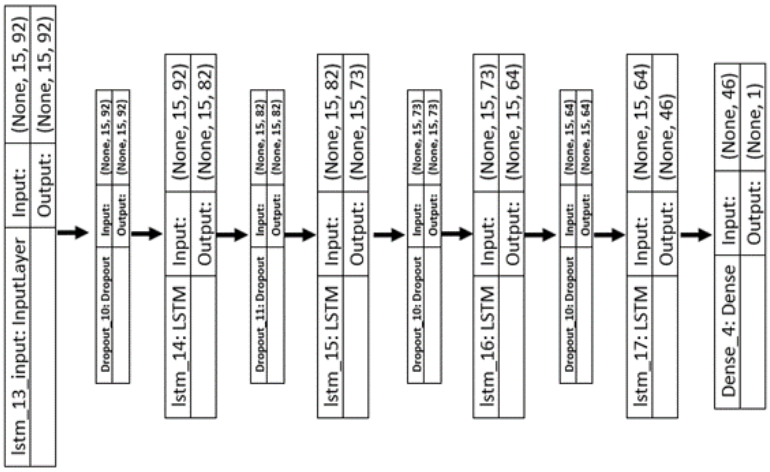
Basically, the technical analysis allows us to understand the price actions such as support, resistance, or trends. Moreover, several indicators can be extracted from the historical prices to support our future price forecasting such as oscillators, measures of volatility, and the most recognized indicator and method is the moving averages. Finally, the last category can be sub-divided into mainly two classes of methods; 1) Linear Models, 2) Non Linear Models.

There are many time series analysis models based on an autoregressive method that has successfully been adopted for stock price forecasting such as autoregressive integrated moving average (ARIMA) [7]. The essential issue with these methods that they lack the inclusion of latent dynamics existing in the data. Usually, the interdependencies among the various stocks that made the index are not captured by these models. Furthermore, these methods are series-oriented which means the model identified for a series will not fit for the other. Non-linear models involve methods like Generalized Autoregressive Conditional Heteroskedasticity (GARCH) [8], machine learning [9], Deep learning algorithms [10].

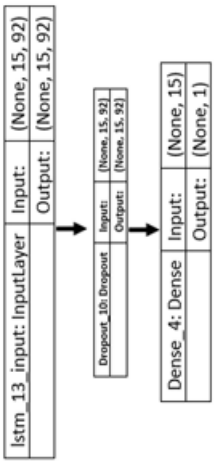
III. Related Work

There are several studies that have been conducted to forecast the stock prices either index or for separate companies. A large number of studies were conducted using linear models such as autoregression moving averages (ARMA). Usually, such models lack accounting for latent factors that are negatively effecting to capture the interactions among different stocks [10]. A non-linear model is, also, studied with various feature-space settings. For instance, Aslam et al. in [11] developed an Islamic securities index forecasting model using artificial neural networks (ANNs) for predicting the performance of the KMI-30 index in Pakistan. In their work, 25 indicators were extracted to train the models and only five of them were identified as the most important factors. Some latent factors were approached in terms of people's sentiments as in [12]. Alamro et al. in [12] studied the effect of news tone and social media attention for Saudi Arabia's stock market index prediction using several multivariate models. The effect of oil prices was, also examined as in [13]. Alotaibi et al. in [13] investigated the use of ANN to forecast Saudi market movement using the Saudi stock market and historical oil prices together. The market movement detection is a reformation of the market prices prediction (continuous value) to a categorical response (i.e., classification) where the target is to predict the market direction to either move up or down by time. Minqi, et al. in [14] studied several machine learning algorithms using a set of combined technical and macroeconomic indicators to identify the direction of the three major stock price indices in the United States, namely the S&P500, Dow 30, and Nasdaq. Similarly, Jigar, et al. in [15] proposed Trend Deterministic Data preparation layer to discretize the Indian stock index prices into trends. Then, they studied several machine learning algorithms such as support vector machines (SVM) Random Forest, and ANNs. The latent variables of the stock index prices may not be easily captured by finding relationships between the stock price index and other markets such as oil [6] or people's opinions [12]. Jarrah et al. in [16] proposed discreet wavelet transformation (DWT) to reduce noise that might be produced due to latent variables to forecast Saudi stock price trends of a few selected companies. Their study employed a recurring neural network (RNN) to forecast the next 7 days of closing prices pertaining to the chosen sample of companies. Moreover, Zhang et al. in [17] proposed an integration between support vector regressor and fuzzy inference systems to improve prediction performance and to reduce the effect of uncertainty in price prediction due to external factors.

There are few studies that have attempted to test the Saudi market index using deep Learning algorithms [9], [18], [19]. The aim of this study is to forecast the Saudi Stock Market Index by integrating Bloomberg technical indicators and machine learning. A long short-time memory network(LSTM)



(a) LSTM 01 architecture



(b) LSTM 02 architecture

Fig. 2: Long Short Time Memory Deep Learning Model Architectures

integrated with the Least shrinkage and selection operator (LASSO) regression to predict midterm TASI prices (see Section IV.

IV. Method

In this section, we discuss the proposed framework and explain the adopted indicators for forecasting midterm TASI index prices. In Figure 1, we illustrate the main components of the proposed framework. It consists of three main stages: data preparation, modeling, and deployment stages. In the Data stage, a target market to be analyzed should be accessed to retrieve historical data with the longest possible horizon. This retrieved raw data, then, is analyzed to extract features using Bloomberg indicators that are commonly used by experts. Finally, the extracted features may require some cleaning and imputation as some values may not be defined correctly (i.e., NAN values). Therefore, the clean and imputation can conducted per feature-column such as MinMax Normalization, and the mode or average values depending on the nature of the target feature. Once, the data is cleaned and prepared, Modeling stage starts by using a traditional machine learning method LASSO to indicate the importance of each feature. As a result, some features will be selected for modeling the machine learning algorithms. In, the deep-learning modeling, the framework aims to got enough data to build robust models for TASI index price forecasting. Finally, the trained model is deployed to forecast a midterm index price from the testing data.

A. Feature Extraction

The features considered in this study can be categorized into five types: binary patterns, Bloomberg's Elliott wave, simple moving average, trend-based, and Oscillator-based. Technical indicators offer a wider vision to which price action can be analyzed. Usually, experts use the technical indicators to serve their three main objectives: to alert, to confirm, and/or to predict. In this study, we are trying to evaluate and investigate Bloomberg's most common indicators for predicting index prices using deep learning algorithms. In the following, we briefly describe each indicator category and list all used indicators in Table I.

1) Binary Patterns is the forecast of the index price movements. There are six indicators (see Table 1) that are calculated by reducing price fluctuations to a series of patterns consisting of up/down moves. Then, it compares the current patterns of up/down moves to similar historical patterns. Usually, these indicators are used to predict the most recent pattern moves.
2) Bloomberg's Elliott Wave indicators are insights into macro-level price trends. The Elliott Wave is used to produce waves and wave counts displayed on the price chart for the index price. The indicators include several measures which apply a statistical approach based on Bayesian probability to identify data waves. This allows the calculation and presentation of up to three possible wave counts, each with its own inherited probability. Moreover, the Elliott Wave may display defaults to the "primary" or the most probable wave count. To aid in analysis, the major level wave count can be augmented with up to two levels of sub-waves (intermediate and minor).
3) Simple Moving average, a simple, or arithmetic is calculated by adding the closing price of the index for several time periods (window lengths) then dividing this total by the number of time periods. For example, a simple moving average of five days window is calculated as:

$$SMAVG(5) = \frac{1}{5} \times \sum_{i=1}^{5} price_i \tag{1}$$

4) Trend – based
   - Bloomberg Trender Indicator is an adaptive indicator that attempts to capture the majority of the position profit while minimizing whipsaws. The indicator reflects a stay just out of range of the typical pullbacks in price within the trend. For this purpose, two indicators are used TrndrUp and TrndrUp in this work.
   - TrendStall: It points at which a trend is losing momentum and is likely to stall or consolidate. The determination is based on a Rate of Change of the ADX. The parameters for the ADX, Rate of Change, and Moving Average of the Rate of Change are user adjustable.
5) Oscillator indicators:
   - Chameleon Oscillator (SASEIDX) calculates how many overbought/oversold criteria are being met for three momentum indicators, Bollinger Bands, RSI, and Stochastics. The indicator value varies from -6 to 6.
   - Fear/Greed indicator measures the ratio of buying strength to selling strength. It shows whether the Bulls or the Bears are in control at a particular point in time. It is an oscillator based on the on-balance calculation of the true range. It is an excellent oscillator for divergence analysis and for identifying trend persistence.

TABLE I: Bloomberg Indicators used as Features to build the DL models

Category Indicators	Indicator (Features)	Code
Binary Patterns	Delta Line	F1
	Delta Minus	F2
	Delta Plus	F3
	Probability (SASEIDX)	F4
	Signal Aggregate (SASEIDX)	F5
Bloomberg's Elliott Wave	Intermediate Wave	F6
	Major Wave	F7
	Minor Wave	F8
Oscillator-based	Chameleon Oscillator (SASEIDX)	F9
	FG(5) (SASEIDX)	F10
Simple Moving average	SMAVG (5) on Close	F11
Trend – based	ADX MA(5)	F12
	ADX ROC(5)	F13
	TrndrDn	F14
	TrndrUp	F15

B. Machine Learning Methods

The proposed method employed LASSO regression and LSTM. The LASSO has been used for feature selection. Its main role is to retain to identify the important features from the Bloomberg 15 factors while demolishing others that are not contributing to forecasting of TASI prices. Once the important features are determined, the Long short term memory network is adopted to make the final price predictions. LSTM is a special type of recurrent neural networks (RNNs) where neurons are connected with feedback loops to allow communication of data among neurons in a forward and backward layers. In this work, compared two LSTM architecture shallow versus deeper one. Figure 2 shows the two architectures. Other traditional machine learning algorithms has been adopted. Support vector regression (SVR), Decision Trees, Random Forest Regressor, and Linear Regression have been tested in this work.

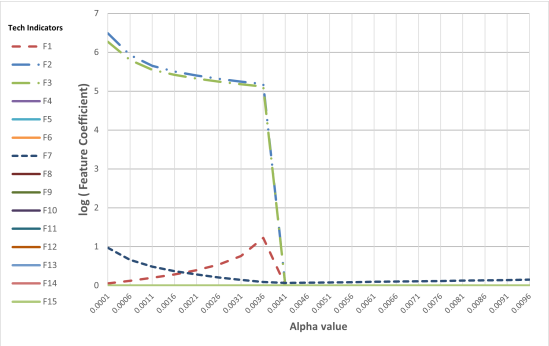


Fig. 3: Feature Selection: F1, F2, F3, and F7

V. Experimental Work

- **Dataset Description** The collected data represents the daily closing prices of the Tadawul All Share Index (TASI). The close price is set as the target variable in this study. The time span of the data is between Jan 1994 to October 2021 with total observation points of 6545. The has been collected from Bloomberg which is the most trusted source of market data worldwide. In addition, TASI data includes also the open, high, and low prices of the index as part of the input of analysis. This dataset represents the entire life of the index from its inception till today. In total, there are 18 factors where 15 of them are popular technical indicators used by Bloomberg users. The reset three are Open, High, and Low prices. Table I lists the 15 factors as categorized into five groups.
- **Models settings:** A LSTM model is developed in this work to forecast TASI prices over a midterm horizon as shown in Figure 5. This model is optimized using the least squared loss function, which represents the continuous loss of prices between the actual and predicted prices. Model LSTM01 structure has five LSTM hidden layers with 92 neurons at the first layer and reduced by 10 percent on each consequent hidden-layers. Dropout layers are introduced among the LSTM layers to reduce the chances of over-fitting during the model training phase. Furthermore, a mini batch is adopted to speed up the training process and mitigates the issue of improper weights initialization. Finally, another model shallower LSTM model is created (LSTM02). The LSTM02 has only 1 hidden layer with 92 neurons.
- **Results and Discussion** Firstly, the data has been prepared for ML experiments. There are several missing information due to the extracted technical analysis methods. Therefore, mode and average imputation are used to fill in missing information. In feature selection, the Lasso method is adopted to rank them using associated

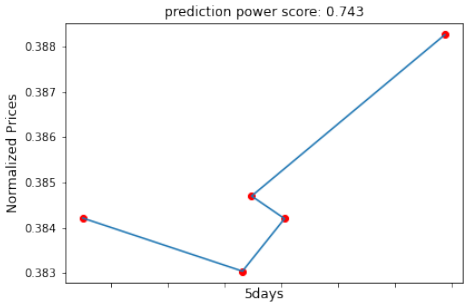


Fig. 4: LSTM01: Five days Forecasting with predictive power: 74.3%

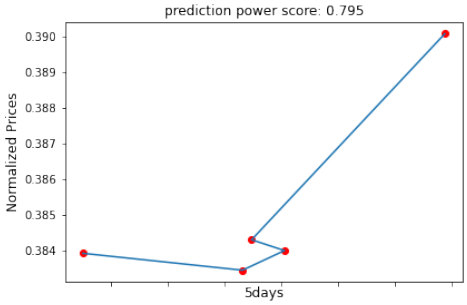


Fig. 5: LSTM02: Five days Forecasting with predictive power: 79.5%

coefficients. Figure 3 shows the effect of increasing the alpha parameter in the Lasso method. Some of the features have zero coefficient at small alpha. This means they contribute less in their linear relationship with the response variable (Closing price). Hence, we can remove them from the modeling stage. Only four features (F1, F2, F3, and F7) were found significant to be considered in modeling the close prices. Training and batch sizes are found empirically. Figure 7 shows the TASI forecasting for 10 months (Jan to October 2021) using different parameter settings. It is obvious that 100 to 150 epochs with a batch of 5 samples for faster convergence is suitable to build the LSTM models.

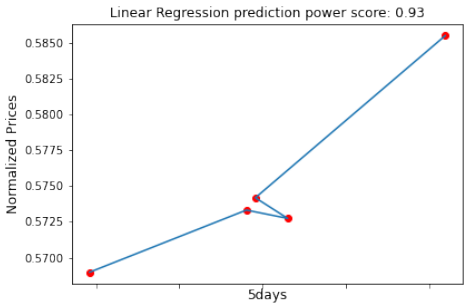
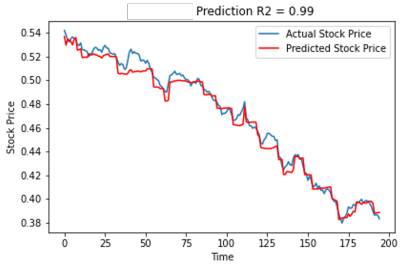
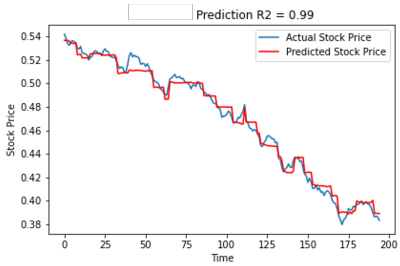


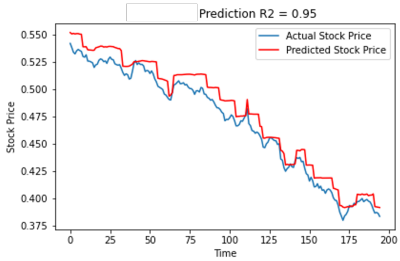
Fig. 6: Linear regression: Five days Forecasting with predictive power: 93%



(a) $epoch = 100, batch = 5$



(b) $epoch = 150, batch = 5$



(c) $epoch = 200, batch = 5$

Fig. 7: Finding best parameters of the LSTM models

Once the LSTM models are configured, two LSTM models have been created using a different number of features. The time span used for creating the machine learning models are the full batch 26 years, 10 years, 5 years up to 1 year. For all these training batches, we set the testing to be the last ten months (Jan. to October 2021). Moreover, 5 days (i.e., weekly) index price prediction is computed using each model. Figure 6 shows the prediction of the first week of the year 2021 using the Linear Regression algorithm. The prediction power of the methods reached 93% compared with 77% for LSTM models (Figure 5. We compared deeper LSTM versus a shallower LSTM designs. In Figure 8, the performance of the LSTM models are depicted in as Bar-chart against the Mean squared error (MSE) of the models. It is clear that the LSTM02 out perform LSTM01 in the first three experiments. This means the shallower models (i.e., simple) methods can be more effective than those complicated ones. This is due to the amount of data/features used to train the models. Also, it can be noticed that as the number of features increases, the deeper model perform better. It is clearer that using all features LSTM01 performed better than LSTM02.

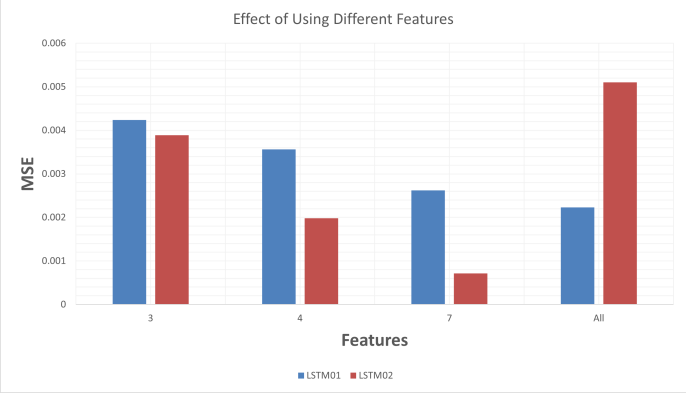


Fig. 8: LSTM Performance: LSTM01 vs. LSTM02

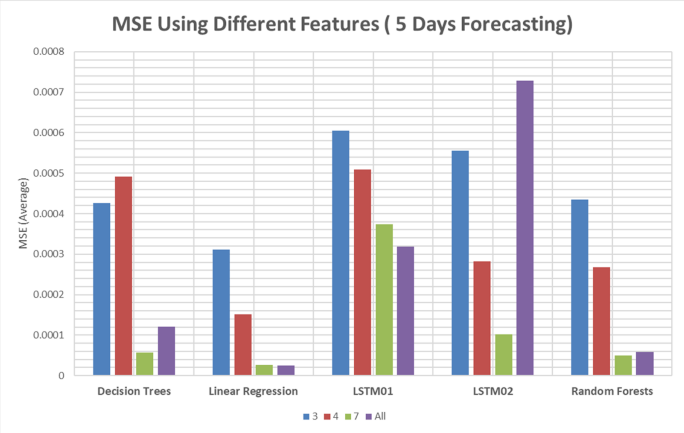


Fig. 9: The effect of using the technical indicators (features) on each model

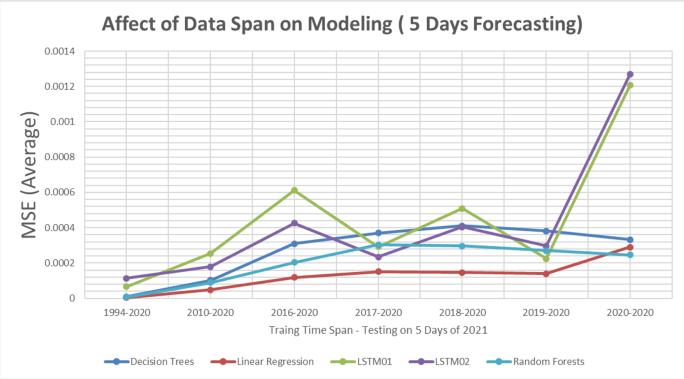


Fig. 10: The effect of using the technical indicators (features) on each model

Usually, the deep-learning models require huge data to learn and construct better models. In another observation, Lasso has effectively selected the most important features for machine learning modeling. By modeling with four features, we observe in Figure 8, the computed mean squared error (MSE) is less than using all features. It is also observed that the most important feature is the F7 (moving average). Therefore, we added the raw three features (Open, High, Low) to the modeling, which produced the best forecasting results (see Figure 8.

Finally, the five machine learning models are compared (see Figure 9). We can notice the Superior performance of the Liner Regression algorithm using all different combinations of features. The Linear regression model's MSE was not dropped significantly when it was trained with 18 features compared to its performance using 7 features. It is also, observed that using complicated models such as LSTM01 will not help in better forecasting. Instead, it may be negatively affecting our system's prediction as not enough examples or features are presented to the model. It is shown in Figure 8 that LSTM01 performance improved by using all features in comparison with other experiments with fewer features.

It is also very evident as the ML models considers larger data, the models perform better in testing. Figure 10 illustrates that effect of using large data for training. All most all models have low testing-MSE when it was trained using the whole 26 years of TASI records. It is observed that deep-learning methods have negatively affected the most when the examples on training stage reduced to a year.

VI. Conclusion

In this paper, we have investigated several algorithms to predicting of stock index values on a weekly forecast horizon. Using the daily historical data of TASI index values during the period Jan 26, 1994, till October 26, 2021, Bloomberg's most-used technical indicators were used and analyzed in this study. Only four were found to be the most significant factors for multi-variate forecasting using machine learning methods. Simple models seem to be more effective in predicting TASI index prices in comparison to more complicated methods. We notice that the shallower LSTM model was Superior to the deeper LSTM design. With this context, Linear regression shows the lowest possible error in predicting TASI index prices compared to all other methods. In the future, more experiments will be conducted to estimate the scalability of simple methods while increasing the data size, and whether complicated machine learning methods will be required to reduce the forecasting errors.

Acknowledgment

The authors would like to acknowledge the help and support provided KFUPM University through funding the project number INFE2120.

References

[1] J.-J. Wang, J.-Z. Wang, Z.-G. Zhang, and S.-P. Guo, "Stock index forecasting based on a hybrid model," *Omega*, vol. 40, no. 6, pp. 758–766, 2012.

[2] Y. Lin, Y. Yan, J. Xu, Y. Liao, and F. Ma, "Forecasting stock index price using the ceemdan-lstm model," *The North American Journal of Economics and Finance*, vol. 57, p. 101421, 2021.

[3] J. Jouini, "Return and volatility interaction between oil prices and stock markets in saudi arabia," *Journal of Policy Modeling*, vol. 35, no. 6, pp. 1124–1144, 2013.

[4] L. Kalyanaraman and B. Tuwajri, "Macroeconomic forces and stock prices: Some empirical evidence from saudi arabia," *International journal of financial research*, vol. 5, no. 1, 2014.

[5] "The Financial Brand artificial intelligence and the banking industry's usd 1 trillion opportunity," https://thefinancialbrand.com/72653/artificial-intelligence-trends-banking-industry/, accessed: 2010-09-30.

[6] A. V. Devadoss and T. A. A. Ligori, "Forecasting of stock prices using multi layer perceptron," *International journal of computing algorithm*, vol. 2, no. 1, pp. 440–449, 2013.

[7] H. F. Assous, N. Al-Rousan, D. Al-Najjar, and H. Al-Najjar, "Can international market indices estimate tasi's movements? the arima model," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 6, no. 2, p. 27, 2020.

[8] J. A. Alzyadat, A. A. Abuhommous, and H. Alqaralleh, "Testing the conditional volatility of saudi arabia stock market: Symmetric and asymmetric autoregressive conditional heteroskedasticity (garch) approach," *Academy of Accounting and Financial Studies Journal*, vol. 25, no. 2, pp. 1–9, 2021.

[9] R. Alamro, A. McCarren, and A. Al-Rasheed, "Predicting saudi stock market index by incorporating gdelt using multivariate time series modelling," in *International Conference on Computing*. Springer, 2019, pp. 317–328.

[10] S. Selvin, R. Vinayakumar, E. Gopalakrishnan, V. K. Menon, and K. Soman, "Stock price prediction using lstm, rnn and cnn-sliding window model," in *International conference on advances in computing, communications and informatics (icacci)*. IEEE, 2017, pp. 1643–1647.

[11] F. Aslam, K. S. Mughal, A. Ali, and Y. T. Mohmand, "Forecasting islamic securities index using artificial neural networks: performance evaluation of technical indicators," *Journal of Economic and Administrative Sciences*, 2020.

[12] R. Alamro, A. McCarren, and A. Al-Rasheed, "Predicting saudi stock market index by incorporating gdelt using multivariate time series modelling," in *International Conference on Computing*. Springer, 2019, pp. 317–328.

[13] T. Alotaibi, A. Nazir, R. Alroobaea, M. Alotibi, F. Alsubeai, A. Alghamdi, and T. Alsulimani, "Saudi arabia stock market prediction using neural network," *International Journal on Computer Science and Engineering*, vol. 9, no. 2, pp. 62–70, 2018.

[14] M. Jiang, J. Liu, L. Zhang, and C. Liu, "An improved stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms," *Physica A: Statistical Mechanics and its Applications*, vol. 541, p. 122272, 2020.

[15] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," *Expert systems with applications*, vol. 42, no. 1, pp. 259–268, 2015.

[16] M. Jarrah and N. Salim, "A recurrent neural network and a discrete wavelet transform to predict the saudi stock price trends," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 4, pp. 155–162, 2019.

[17] J. Zhang, L. Li, and W. Chen, "Predicting stock price using two-stage machine learning techniques," *Computational Economics*, vol. 57, no. 4, pp. 1237–1261, 2021.

[18] T. Alotaibi, A. Nazir, R. Alroobaea, M. Alotibi, F. Alsubeai, A. Alghamdi, and T. Alsulimani, "Saudi arabia stock market prediction using neural network," *International Journal on Computer Science and Engineering*, vol. 9, no. 2, pp. 62–70, 2018.

[19] M. Jarrah and N. Salim, "A recurrent neural network and a discrete wavelet transform to predict the saudi stock price trends," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 4, pp. 155–162, 2019.