# 1. Background [5]

## 1.1 Problem Statement

The goal of this project is to develop an AI-powered application that provides accurate and reliable forecasts of stock prices and movements. The application should utilize machine learning algorithms and historical stock data to analyze patterns and trends, allowing users to make informed investment decisions.

## 1.2 Research pertaining to problem statement

A stock market is a marketplace accessible to the public for trading publicly listed company shares, also known as equities, which represent ownership in the company. The exchange serves as a middleman, facilitating the buying and selling of shares. Stock prediction is primarily defined as the attempt to determine the cost of the stock and provide a sound basis for investors to understand and forecast the future stock price and ultimately return on investments (Akhtar *et al.*, 2022). Relying on only one dataset is not sufficient and could result in an inaccurate prediction (Akhtar *et al.*, 2022). Therefore, we are utilizing machine learning to assist in stock prediction by analysing numerous datasets in order to forecast the demand and trend of stocks. Nonetheless, predicting the performance of the stock market is a challenging undertaking.

We are aware that great fluctuations in the stock market can be caused by unforeseen circumstances and unexpected events, even events such as terrorist incidents, natural disasters, and pandemics (Aslam *et al.,* 2021:1149). As a result, the program uses advanced algorithms to continuously learn and adjust, guaranteeing that the projections change to reflect shifting market conditions. Our goal is to provide precise information to individual investors so they may make wise decisions, maximize their investment plans, and perhaps improve their financial prospects.

# 2. Choice of Techniques [5]

Why did you finally choose the techniques you opted for? knowledge representation, learning algorithm, search order and operators, etc

## 2.1 Knowledge representation

The code represents the data using pandas DataFrames, which provide a convenient and efficient way to handle and manipulate tabular data. The stock price data is stored in CSV files, which are read and processed using pandas. This is to provide easy and constant access to the data across multiple devices. Training data is received from an API (yahoo finance) to allow for the most accurate and up-to-date training data. We display the results of the model using a graph in order to quickly see how accurate our model's predictions are in comparison to the actual stock prices.

## 2.2 Learning algorithm

Learning algorithm: The program uses a Long Short-Term Memory (LSTM) model with a sequence of 60 time steps and 3 layers. LSTM models are a type of recurrent neural network (RNN) that can capture long-term dependencies in sequential data, making them suitable for time series forecasting tasks. Thus, the choice of the LSTM model is appropriate for this type of problem because it can learn long-term dependencies and handle time-series data well. We also used a sequence of 60 time steps as it allows the model to learn patterns over a longer time horizon, but not too long as it requires a lot of memory.

The data is normalized using a MinMaxScaler from scikit-learn, this is a good way to normalize the data since it scales the data to a range between 0 and 1, which is suitable for neural network models. It also helps improve the convergence and stability of the LSTM model during training. The scaler is saved for later use during prediction to transform the test data in the same way.

We also made use of the Adam optimizer with mean absolute error loss, this is because it is a widely used optimization algorithm that works well with deep learning models. The mean absolute error is a good choice of loss function because it is a standard metric for regression problems, displaying a good measure of how confident our model is of its predictions.

## 2.3 Search order and operators, etc.

There is no explicit search order or operators used in our model. The training process involves preparing the input sequences, defining the LSTM model architecture, compiling the model with an optimizer and loss function, and fitting the model to the training data and thus has no need for a search order method or operators. However, within the LSTM model architecture, dropout layers are used as a regularization technique to prevent overfitting. Dropout randomly sets a fraction of input units to 0 during training, which helps in reducing the model's reliance on specific input features and promotes generalization.

# 3. Modeling, Simulation Experiment and Program Design [15]

Give a brief description of your models, simulation experiment, parameter setting and program design.

## 3.1 Brief description of models

Various libraries and modules are used to make up the LSTM model.

* Keras: Is a neural network library that acts as an interface for the TensorFlow library.
* TensorFlow: Is a library for machine learning and artificial intelligence, with a focus on neural networks.
* keras.model.Sequential: Is a recurrent neural network model that facilitates the stacking and ordering of layers. It is ideally suited for dealing with sequential data, making it very applicable for the time series data format of stock data.
* keras.model.LSTM: Is the LSTM layer where the learning takes place. It has the following parameters:
  + units: an integer value that describes the dimension of the output space of the LSTM layer. In other words, it determines how many datapoints are going into the next layer.
  + return\_sequences: Determines whether the output of the layer will be kept as an array of sequential data (when set as ‘true’) or not (when omitted or set as ‘false’).
  + input\_shape: Defines the shape/dimension of the input data. In our case our input into the initial LSTM layer is a variable length array (depending on the size of the training data set) of timestep-sized arrays, which each contain a single element.
* keras.model.Dropout: Determines the factor of the amount of neurons in the model that will be randomly turned off/ignored. It helps to avoid overfitting data and ending up with a model that is suited only for a specific situation, and not applicable to any other situations.
* keras.model.Dense: Sets the dimension of the output layer. Setting it to the value of one combines the output into a single value.
* optimizer: An optimizer is an algorithm that influences how the model assigns weights to datapoints and rate at which it learns (how quickly it can reduce the losses).
* loss: Determines how the model calculates the error between its own predictions and the true values for the training data provided to it during the fitting step.

Our final model is layered as follows:

model = Sequential()

model.add(LSTM(units=50,return\_sequences=True,input\_shape=(X\_train.shape[1], 1)))

model.add(LSTM(units=50,return\_sequences=True))

model.add(Dropout(0.2))

model.add(LSTM(units=50))

model.add(Dropout(0.2))

model.add(Dense(units=1))

model.compile(optimizer='adam',loss='mean\_absolute\_error')

## 3.2 Simulation experiment

For our simulations, an LSTM model is provided with a dataset containing the historical opening stock price values of a given stock (the data has been normalized and arranged into timesteps). The model trains itself on this dataset, and outputs its loss value for each training epoch (iteration). It is then provided with another dataset, containing the historical opening stock price values of a given stock (also normalized and arranged into timesteps). The model does predictions on this testing dataset, providing a corresponding predicted value for every stock value. The normalization on the predicted values is reversed. The actual stock prices and the corresponding predicted prices are visually displayed by plotting them both on the same graph. As a measure of accuracy, both the root mean squared error and the mean absolute percentage error are calculated between the actual and predicted stock prices.

## 3.3 Parameter setting

The following parameters were used in our final model:

* Model: 3 LSTM (units = 50) layers and 2 Dropout (0.2) layers, which we defined as (3, 2)
* Model optimizer: adam
* Model loss: mean absolute error
* Timesteps: 60
* Epochs: 200
* Training and testing on the same stock, by splitting the dataset up into a 0.75/0.25 training/testing dataset length ratio.

## 3.4 Program design

# 4. Training/Testing Strategy [10]

Testing and experiments used to arrive at your final strategy? Detail of experimental setting and strategies, improvement and hybrids (if any).

## 4.1 Testing and experiments used

The initial layering and parameters, as per the most widely used configuration found whilst conducting research for the model, were set up in the following way:

* Model: 4 LSTM (units = 50) layers and 4 Dropout (0.2) layers, which we defined as (4, 4)
* Model optimizer: adam
* Model loss: mean squared error
* Timesteps: 60
* Epochs: 100
* Training and testing on different stock types

We then started assessing the impact of each parameter by adjusting it while holding the parameters constant. For each adjusted value of the parameter being inspected, the training and prediction was repeated six times. As a measure of the effect that a parameter has on the performance of the model, the following values were recorded to be used as metrics:

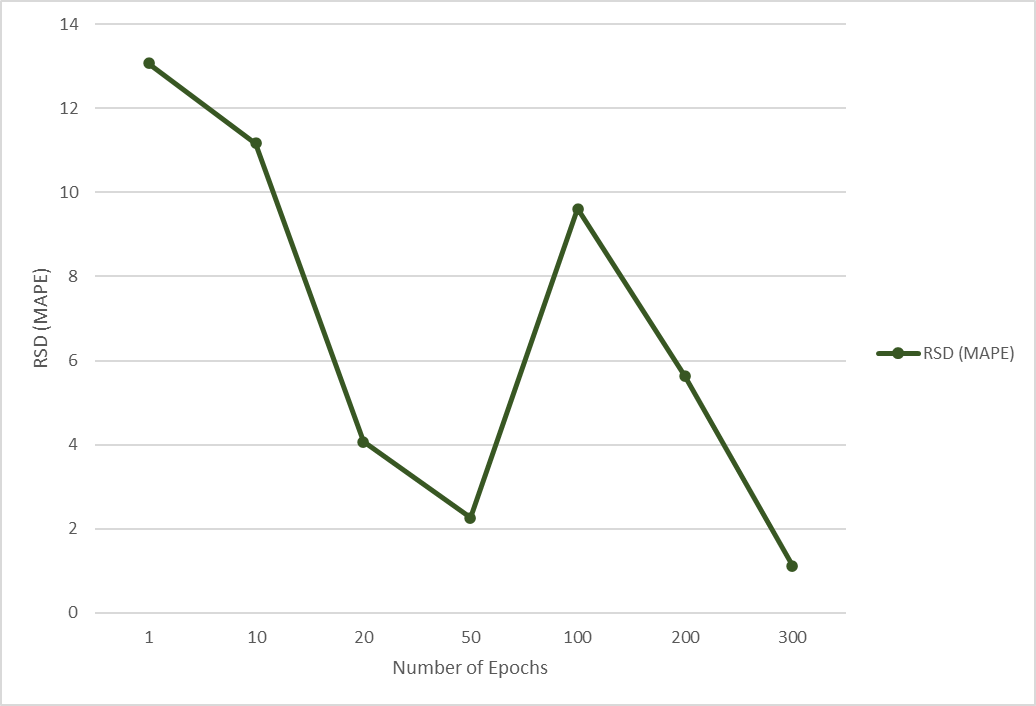
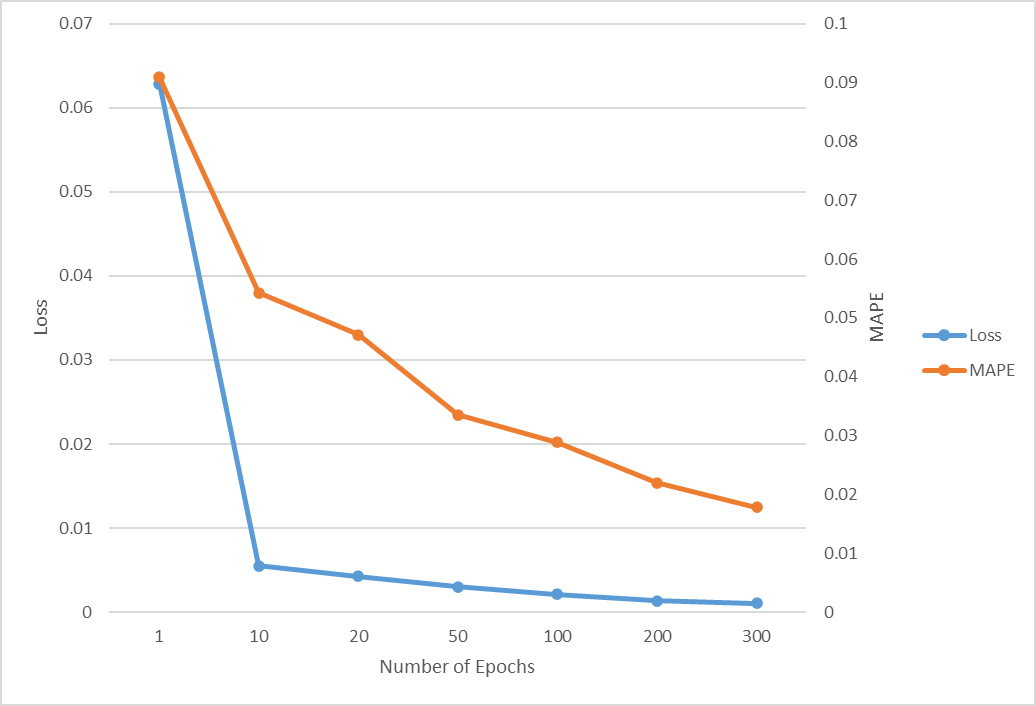
* The loss after the final epoch when training the model.
* The root mean squared error (RMSE) between the actual and the predicted stock prices.
* The mean absolute percentage error (MAPE) between the actual and the predicted stock prices.
* The relative standard deviation of each of the six repeated runs.

Lower values are considered more desirable for all four metrics. The values for the metrics, and the graphs for all the experimental runs can be found in the “Experimental Results.xlsx” file among the supplementary submission files.

## 4.2 Detail of experimental setting and strategies

## 4.2.1 Epochs

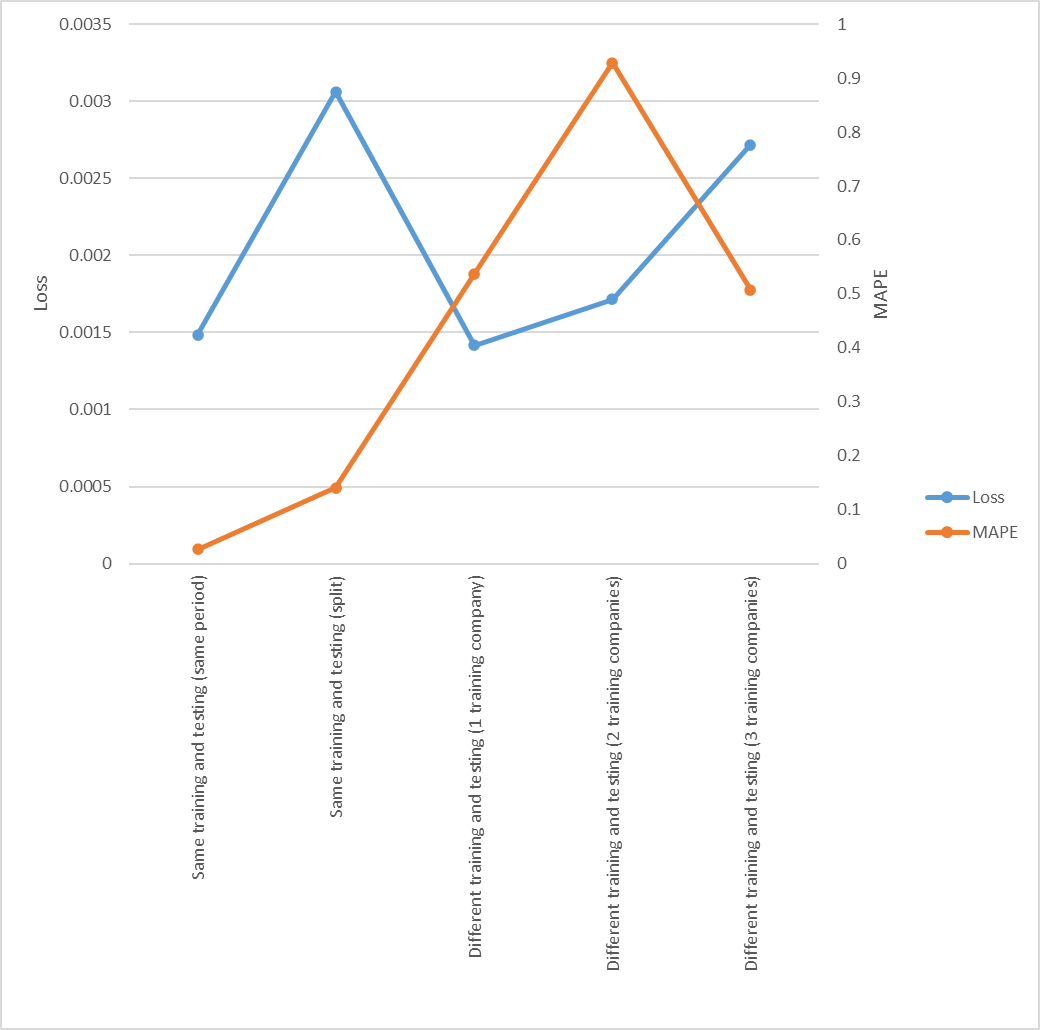
The first parameter to be assessed was the number of epochs, in other words, the number of times that the model iterates through the training dataset while it is being taught. The following values were tested: 1, 10, 20, 50, 100, 200, and 300.



Based on the results, it is apparent that the loss, MAPE, and RSD improves with a higher number of epochs. However, computation time also increases with the number of epochs. Since there seems to be a diminishing return on the improvement of the metrics with an increase of the number of epochs, we have selected the value of 200 to be used for the final model.

## 4.2.2 Datasets

Different combinations of datasets used for training and testing were tested. The first was to train and test on the same stock dataset, by spitting the dataset up into training and testing sections. Next was training on one stock dataset, whilst testing on a different stock dataset. A similar test was done, except using two different stocks for training, and one for testing. Lastly, three different stock datasets were used for training, and one for testing. For interest sake, a test was also done on training and testing on the same data. In all the cases, the same amount of datapoints were used for training, regardless of the dataset combination.



As expected, training and testing on the exact same data yields the best results. However, this would not be considered correct practice, and was merely done in the interest of confirming the outcome. Otherwise, across all metrics, training and testing on the same stock by splitting up the dataset into a training and testing portion, yielded the best results. Therefor it has been chosen as the approach for our final model.

## 4.2.3 Dataset Length

The effect length of the datasets used for training was also assessed. Despite not being a fixed parameter for the model, it was assessed to get a better understanding of its effect on the predictions, as well as to provide a guideline length to use for datasets. We looked at using 6 months, 1 year, 3 years, and 5 years of training data.

Looking at the results, it is clear that in terms of accuracy (MAPE), using 3 and 5 years perform better. The variance (RSD) amongst the individual runs seems to increase with increasing training data length. A possible explanation could be that the effect of increasing long-term dependencies introduces more inconsistency whilst the model is being trained, with it not being able to assign the same importance to certain trends that happened long ago relative to the dataset length every time that it trains the model.

We decided that 3 years of training data seems to provide the best combination of accuracy and consistency as well as computation time, since longer training sets result in longer runs.

## 4.2.4 Timestep Size

The timestep size determines the length of the sequence of values that the model uses for every prediction that it makes. We looked at timestep sizes of 10, 30, 60, and 120. From the results, only a marginal increase in accuracy is seen by increasing the timestep size to 60 and then further to 120. The value for the loss also goes up when using 120 timesteps, indicating that it can still benefit from increasing the epoch value. The variance for 60 and 120 timesteps are very comparable. We decided to go with a timestep size of 60, since it performs just as well as 120 timesteps at 200 epochs, with the added benefit of having shorter runs.

## 4.2.5 Model Optimizer

We looked at the following four model optimizers: Adam, Stochastic Gradient Descent (SGD), Nadam, and Root Mean Square Propagation (RMSprop). All the optimizers performed comparably, with the exception of SGD, which performed significantly worse than the others across all metrics. Since Adam seems to be the most widely used optimizer among the models considered whilst conducting research for the model, we opted to go for it for the final model.

## 4.2.6 Model Loss

We looked at the loss – that is how the model calculates the error between its own predictions and the true values for the training data provided to it during the fitting step. Four different techniques were used, Mean Squared Error (MSE), Mean Absolute Error (MAE), Huber Loss, and Mean Squared Logarithmic Error (MSLE). MSE and MAE performed marginally better than the other two in terms of accuracy, with MAE also outperforming MSE in terms of variance. Based on that, we decided to go use MAE for the final model. We did, however, notice an increase in the loss for MAE, indicating that it can benefit further from increasing the number of epochs.

## 4.2.7 Model Layering

For model layering, we looked at different layering combinations in terms of how many LSTM, and Dropout layers to add. We made use of a naming convention for different layered models as (L, D), where L is the amount of LSTM layers (with units = 50), and D is the amount of Dropout layers (0.2) present in the model. We looked at the following differently layered combinations: (4,4), (4,0), (2,4), (2,0), (8,4), (4,8), (8,8), (3,2). From the results, we determined that the (3,2) layered model yields the best combination of accuracy and variance.

## 4.3 Improvements

## 4.4 Hybrids

# References

Aslam, S., Rasool, A., Jiang, Q. and Qu, Q., 2021, July. LSTM based model for real-time stock market prediction on unexpected incidents. In *2021 IEEE International Conference on Real-time Computing and Robotics (RCAR)* (pp. 1149-1153). IEEE.

Akhtar, Md.M., Zamani, A.S., Khan, S., Shatat, A.S.A., Dilshad, S. and Samdani, F. 2022. Stock market prediction based on statistical data using machine learning algorithms. *Journal of King Saud University - Science*, 34(4), 101940. <https://doi.org/10.1016/j.jksus.2022.101940>.

Non-academic resources:

Resources used in the construction of our application:

YouTube Video - "Machine Learning & Predictive Modeling in Finance" by CodeEmporium (2020). [Online]. Available: https://youtu.be/QIUxPv5PJOY. [Accessed: 29 March 2023].

YouTube Video - "Stock Prediction Using Machine Learning" by AlgoTrading101 (2020). [Online]. Available: https://youtu.be/KYc0EFN-VnM. [Accessed: 29 March 2023].

YouTube Video - "Stock Price Prediction using Machine Learning" by Dataquest (n.d.). [Online]. Available: https://www.youtube.com/watch?v=1O\_BenficgE&ab\_channel=Dataquest. [Accessed: 29 March 2023].

Simplilearn. (n.d.). "Stock Price Prediction using Machine Learning." [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/stock-price-prediction-using-machine-learning. [Accessed: 29 March 2023].

ProjectPro. (n.d.). "Stock Price Prediction using Machine Learning Project." [Online]. Available: https://www.projectpro.io/article/stock-price-prediction-using-machine-learning-project/571. [Accessed: 29 March 2023]

YouTube Video - "Stock Market Prediction Using Machine Learning" by Krish Naik (2019). [Online]. Available: https://www.youtube.com/watch?v=6bzkh\_sOCoE. [Accessed: 29 March 2023].

ProjectPro. (n.d.). "Stock Price Prediction using Machine Learning Project." [Online]. Available: https://www.projectpro.io/article/stock-price-prediction-using-machine-learning-project/571. [Accessed: 29 March 2023].

Dataconomy. (2023, January). "Stock Prediction with Machine Learning." [Online]. Available: https://dataconomy.com/2023/01/stock-prediction-machine-learning/. [Accessed: 29 March 2023].

Towards Data Science. (n.d.). "Predicting Stock Prices Using a Keras LSTM Model" by Aman Dalmia (n.d.). [Online]. Available: https://towardsdatascience.com/predicting-stock-prices-using-a-keras-lstm-model-4225457f0233. [Accessed: 29 March 2023].

GeeksforGeeks. (n.d.). "Stock Price Prediction using Machine Learning in Python" by Ishan Garg (n.d.). [Online]. Available: https://www.geeksforgeeks.org/stock-price-prediction-using-machine-learning-in-python/. [Accessed: 29 March 2023].