Department of Computer Science

MSc Data Science and Analytics

Academic Year 2022-2023

Dissertation Title:
Predicting the Indian Stock Market using NIFTY-50 Index

Name: Drishti Harshit Doshi

Student ID: 2203898

A report submitted in partial fulfillment of the requirement for the degree of Master of Science

Brunel University
Department of Computer Science
Uxbridge, Middlesex UB8 3PH
United Kingdom
Tel: +44 (0) 1895 203397
Fax: +44 (0) 1895 251686

Abstract:

The stock market is a non-linear, non-parametric, dynamic, and highly volatile sector of the financial market. Stock market prediction is a fascinating area of research which has replaced traditional trading systems to high-frequency automated algorithmic trading systems. In the Indian Stock Market, there are a total of 50 companies listed in the NIFTY-50 Index which are from 12 different sectors as per NSE (National Stock Exchange of India). This research study aimed to perform a comparative analysis of these 50 stock companies listed in NIFTY-50 Index using deep learning and machine learning methods which include ANN, CNN, LSTM, Hybrid Network, MLP, DT, RF, SVM, Gradient Boost, and XG Boost. Technical indicators were used in the input data and the 10th day closing price prediction was made. The performance of the model was evaluated using RMSE and MAPE scores. Results show that CNN was the best performing model with RMSE score in the range of. This was followed by XG Boost and Gradient Boost. Best five performing companies were Tata Steel, PowerGrid, ONGC, NTPC, ITC. This study was successful in analyzing stock prediction techniques on the Indian stock market.

Acknowledgements:

I would like to express deep and sincere gratitude to my Supervisor Dr Matloob Khushi, for his endless support and guidance throughout this dissertation project. With the help of his constant motivation and enthusiasm, I was able to understand the domain of the Stock Market and perform the analysis. I would also like to express my gratitude to my family for their constant support throughout the dissertation journey.

Sign in the box below to certify that the work carried out is your own. By signing this box you are certifying that your dissertation is free from plagiarism. Make sure that you are fully aware of the Department guidelines on plagiarism (see the student handbook). The penalties if you are caught are severe. All material from other sources must be properly referenced and direct quotes must appear in quotation marks.

I certify that the work presented in the dissertation is my own unless referenced

Signature: Drishti Doshi

Date: 11/09/2023

Insert a word count. This is the sum of the words in all the chapters only. The sum should exclude the words in the title page, abstract, acknowledgements, table of contents, references and any appendices.

TOTAL NUMBER OF WORDS: 12287

[*All the above should not exceed this one page]*

Table of Contents:

List of Tables:

<u>*List of Figures:*</u>

Abbreviations:

IPO: Initial Public Offerings
GDP: Gross Domestic Product
EMH: Efficient Market Hypothesis
NASDAQ: National Association of Securities Dealers Automated Quotations
NYSE: New York Stock Exchange
S&P 500: Standard and Poor US Stock Exchange
HFT: High Frequency Trading
NSEI: National Stock Exchange, India
NIFTY-50 INDEX: National Stock Exchange Fifty
ANN: Artificial Neural Network
AR: Auto Regression
ARMA: Auto Regressive Moving Average
ARIMA: Auto Regressive Integrated Moving Average
SARIMA: Seasonal Auto Regressive Integrated Moving Average
CNN: Convolutional Neural Network
DL: Deep learning
LSTM: Long Short-Term Memory
ML: Machine Learning
MAPE: Mean Absolute Percentage Error
RMSE: Root Mean Square Error
MLP: Multi-Layer Perceptron
RNN: Recurrent Neural Network
KNN: K-Nearest Neighbour
NN: Neural Networks
RL: Reinforcement Learning
GRU: Gated Recurrent Unit
FTSE: Financial Times Stock Exchange
RF: Random Forest
DQN: Deep Q-Network
DDQN: Double Deep Q-Network
Dueling DDQN: Dueling Double Deep Q-Network
NLP: Natural Language Processing
GSE: Ghana Stock Exchange
PCA: Principal Component Analysis
TA-Lib: Technical Analysis Library
ReLU: Rectified Linear Unit
RBF: Radial Basis Function
ATS: Algorithmic Trading System
TI: Technical Indicators

Chapter 1: Introduction

1. A) Stock Market:

The stock market has always been a captivating force of the economic gradient that gains the attention of the people globally. The stock market is the platform where the shares or stocks of a company are traded. (Hiransha et al., 2018) It has two components i.e., a*) the primary market* is where novel securities are introduced in the market i.e., Initial Public Offerings (IPO) and *b) the secondary market* is where the investors trade with existing securities. (Hiransha et al., 2018) This market is highly influenced by the global situation and the country's economic, political, and company fundamental factors i.e., several underlying micro-economic and macro-economic factors that influence the functioning of the stock investments system.(Table 1) These financial markets play a significant role in the lives of the common man. It helps to diversify investment options such as index funds, mutual funds, and hedge funds. Even the government sectors allocate a portion of their healthcare, employment, and retirement funds into the stock market with the hope of gaining profitable returns and contributing to society. The stock market has a direct influence on the economy of the country. (Masoud, 2013) Hence it conserves a huge amount of attention and energy from people worldwide.

Macro-Economic Variables	*Micro-Economic Variables*
Interest rates	Revenue Growth
GDP	Dividend Policy
Inflation	Supply and Demand Dynamics
Political News	Debt levels
Crude Oil Prices	Company's Management
Exchange Rates	Legal and Regulatory System

Table 1: Macro-Economic and Micro-Economic Factors Affecting Stock Market

1. B) Theories on Stock Market:

Renowned researchers worldwide continue with an inquisitive quest: Can we find a viable way to predict stock prices efficiently? Initially, this faith was criticized massively by the Efficient Market Hypothesis (EMH) and Random Walk Theory. (Weng et al., 2017) The concept of EMH was introduced in the 1960's by Paul A. Samuelson and Eugene F. Fama. It states that the market prices reveal all the necessary information, and it is nearly impossible to outperform the market performance by using prediction techniques. (Lo, 2007) Additionally, it conveys that the historical data is informationally inefficient in predicting stock prices and it is not feasible to derive extra revenues by exploiting price predictability. (Singh and Khushi, 2021) This thought is closely aligned with the Random Walk Theory which states that the stock market is a random walk, and it is a fool's game to predict them. (Shah, Isah and Zulkernine, 2019) On the contrary, people do have faith in Warren Buffet's capabilities to beat the S&P index repeatedly. The EMH's criticism has led to a rise in research that counter-questions its validity. The newer proficient means to analyze the market i.e., by using technical analysis, charts trends, statistics, data mining, artificial intelligence, and econometrics assures that the market can be predicted and provides momentum to financial researchers to explore newer means of stock price predictions.

1. C) Influence of Technology on Stock Predictions:

Advancement in technology has caused a paradigm shift globally. Machine learning methods have been comprehensively used in healthcare, finance, commercials, and politics. (Khushi and Meng, 2019) This revolution in technology has impacted the process of trading globally. Traditionally, buying and selling stocks was done by humans commonly known as traders and investors. In today's time, with Industry 4.0, financiers prefer intelligent trading systems for quicker decisions over conventional fundamental analysis. (Gadgil et al., 2021) In today's era of big data, everything is multiplying manifold including the complexity of a company's functioning. It is going to be difficult in the future to get a clear picture of the world's economy with the ever-rising interference of data and technologies. The rise in the rate of data means a greater need to process and extract knowledge from the data. This has massive implications for the stock market as this market is loaded and flourishes with timely data. However, this data is raw, i.e., just indicators. It does not convey much meaning unless explored with limitless possibilities of data science and artificial intelligence. Stock market predictions are the most fascinating subject for researchers all around the world. It is a demanding domain but at the same time is filled with challenges and loopholes. In the next section, we will explore the domain of financial markets.

1. D) Financial Markets:

Financial markets provide a one-of-a-kind platform for investing and trading. Financial researchers across the world are curious to use machine learning and artificial intelligence for understanding and forecasting stock and forex markets. One of the primary goals of the investor is to predict the stock price in such a manner that he can buy or sell the stock before the actual stock price decreases or increases. (Gadgil et al., 2021) The versatile financial market can be studied and analyzed by two methods, i.e., fundamental analysis which involves reviewing the intrinsic value of the company, the performance of the country's economy, the political situation, etc. Another method is technical analysis which involves understanding the financial market with the help of statistics, and historical data, and identifying trends and patterns in stock charts. (Patel et al., 2015). Machine learning for financial analysis involves extracting the raw data known as input from the historical data available on the websites. The output data is what needs to be predicted. The input and output data are divided into two parts i.e., test and train data. The train data is sent to the machine to find relevant hidden patterns and relationships between the data points. This is followed by using the test data to predict the results. Predicting the stock prices correctly can increase the investor and trader's profit margin. (Parmar et al., 2018) However, this has certain drawbacks as well. Stock market prediction is a time-series analysis. A time series consists of a sequence of observations that are recorded at some specific timestamps. (Wang et al., 2022) It involves a set of data analyzed over time. Time series data can be further divided as per variability with time into stationary and nonstationary data. *Stationary time-series data:* The mean and variance of a time series are constant over time. The time-trend pattern is evident and easily identified. e.g., Heartbeat, timestamps, train. *Nonstationary time-series data* refers to data that constantly change without obvious regularity, they do not have a clear trend and get influenced by various other factors e.g., stock, wind speed, and rainfall. (Wang et al., 2022)

1. E) Stock Predictions: Quest and Problems Associated

The stock market values are non-linear, non-stationary, non-parametric, dynamic, unstructured, and highly volatile i.e., subjected to influence by political situations. Therefore, investors are looking for newer trading methods. Nowadays, Algorithm trading has a wide spectrum of applications from high-frequency trading to portfolio management. NASDAQ and NYSE launched the first electronic trading platform in 1971. From 2010, there was a rapid growth in Algorithm trading. HFT, i.e., Automated high-frequency trading was an integral part of technological advancement. These trading types involve algorithms to make decisions instead of humans. Since then, algorithm trading has been a core of financial analysis adopted at large scale globally. (Grindsted, 2021) Hence, finding an optimal algorithm with the best performance metrics is a challenge in the pursuit of more accurate and profitable stock market predictions. It is essential to predict stock failures and financial distress for the investor who prefers to be safe in the market. (Pattewar, Jain and Kiranmayee,2023)

The National Stock Exchange (India) was ranked 3rd in the world in the equity segment by trade numbers and has been awarded as the world's largest derivatives exchange in 2022. The average daily turnover of NSE was 68 crore Indian Rupees as reported in April 2022. (The Economic Times, NSE Nifty) This research study is done on the Indian stock market - NIFTY 50 index using machine and deep learning methods. Fifty companies under the NIFTY-50 Index have been analyzed for predictions. NIFTY-50 is an index consisting of 50 stocks that represent 12 different sectors of the Indian economy listed on the National Stock Exchange (NSE). (Table 2) This index is widely used as a benchmark for mutual fund portfolios, index funds, institutional investors, and index-based derivatives. (Girish, 2019) India has low financial literacy rates, especially for stock investments. Hence, it is important to educate the public about correct investing strategies. (Gadgil et al., 2021) (Figure 1)

1. F) Dissertation Outline:

The purpose of this project is to use machine learning and deep learning models on the 50 companies under the NIFTY-50 Index to predict the close prices. To design the research process, past works of literature were reviewed to find meaningful answers to the questions below.

- Which dataset to choose for the study? Time period? Source?
- Will the prediction be long-term or short-term?
- What should be the input and output data for prediction?
- Use of additional features i.e., technical indicators for improving model predictability
- How to pre-process the data?
- Which algorithms will be most suitable to use for time-series forecasting?
- What will be the performance metrics used?
- What kind of hyper-parameters to use for optimal tuning of the model?
- How to evaluate the overall outcome of the research project?

In the realm of financial research, there have been a huge number of papers published with the use of multiple algorithms, thousands of variations in data pre-processing techniques, and all studies having a common aim of achieving better model performance. This project involves performing a comparative analysis of various algorithms for predicting the prices of 50 companies under the NIFTY-50 Index (NSEI) and finding the best model. The research study will be guided and supervised by Dr. Matloob Khushi who is a lecturer at the computer science department, at Brunel University London.

Research study topic**:** Predicting the Indian Stock Market using the NIFTY-50 Index

Research Aim: Comparative analysis of various algorithmic models for the NIFTY-50 Index.

To achieve the research aim, the following are the objectives defined.

Research Objectives:

- Pre-processing the NIFTY-50 Indian stock data
- Adding relevant technical indicators for improving model predictability.
- Applying various machine learning and deep learning algorithms.
- Evaluating the model based on performance metrics.
- Optimising the model by using hyper-parameter tuning
- Performing comparative analysis

In this section, we discussed the global importance of financial markets. We highlighted the contradictory arguments of the efficient market hypothesis and the possibilities of forecasting stock prices. Technology has brought a massive impact on the sector of stock trading causing a transition from traditional trading to a highly advanced algorithmic trading system. This research study involves performing a comparative analysis of the Indian stock market using various algorithmic machine and deep learning models. In the next section, we will review the past works of literature and try to understand the advancement of financial research especially in the domain of stock price predictions.

Operator	NSE Indices
Exchanges	National Stock Exchange of India
Constituents	50
Type	Large Cap
Market Cap	US $2.27 Trillion (April 2018)
Weighting Method	Capitalization-weighted

Table 2: NIFTY-50 INDEX

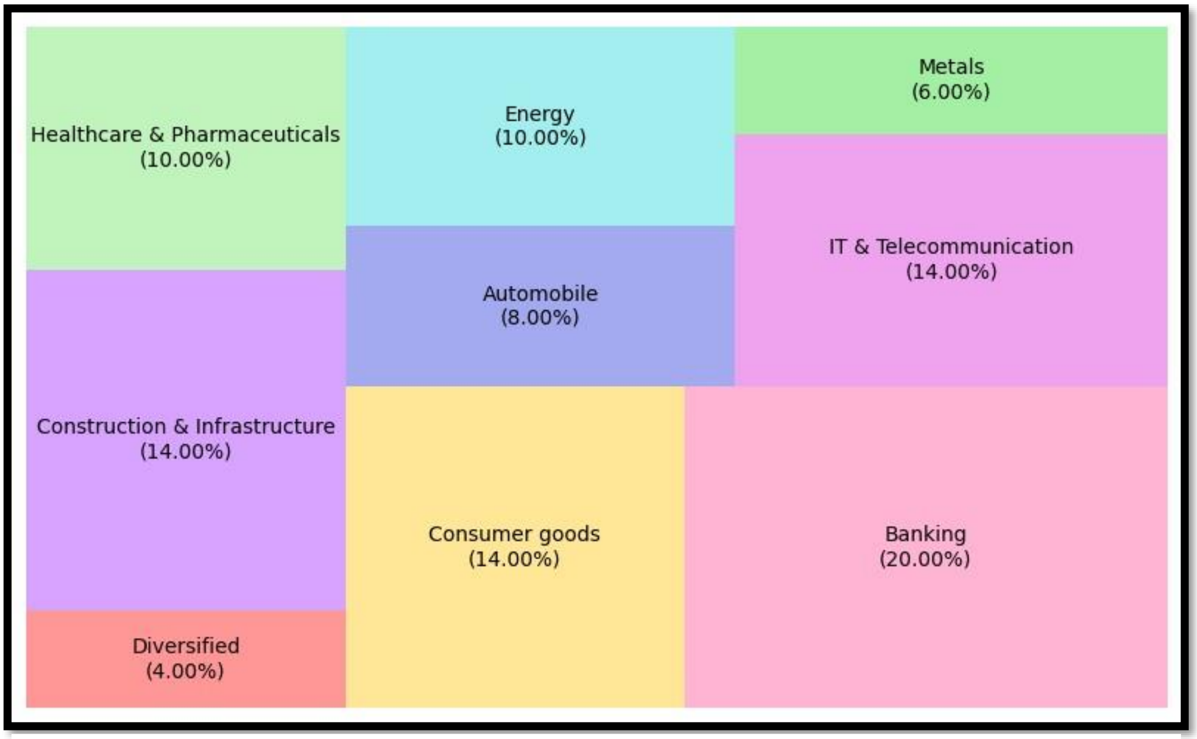


Figure 1: NIFTY-50 Index Stock Company Sectors

Chapter 2: Literature Review

Stock market prediction has been a fascinating area of study for a long span of time. This domain of financial research is flourishing with diverse kinds of literature. In the following section, we intend to explore and discuss the works of researchers by studying relevant literature related to the forecasting of stock prices. Multiple papers were reviewed, out of which only the best papers were selected as per the *3-paper game rule* given in the Research project management module. (CS5704 Week 1 Teaching Materials, 2022) This will help to get a comprehensive overview of the existing knowledge and research gaps in the financial markets. Understanding the literature will contribute to providing an analytical framework and guidelines for carrying out the research study on stock price prediction.

2. A) Literature Review on Stock Predictions using Machine & Deep Learning Methods:

Supervised machine-learning techniques have been widely used in stock price prediction. (Parmar et al., 2018) conducted a study to forecast the price of the financial stocks of a company using ML models. (Kumar, Sarangi and Verma, 2022) conducted a systematic literature review of 30 research papers to answer six research questions on statistical tools, machine learning algorithms, hybrid models, dataset types, performance metrics, and renowned journals. ARIMA was regarded as the most important statistical tool. They found that Neural networks (33%) and ANN (30%) were the most predominantly used algorithms. Hybrid approaches were less commonly used by researchers which accounted for up to 9%. Along with this, it showed the percentage distribution of performance metrics used in the studies, which were as follows: accuracy at 33%, RMSE at 20%, MSE at 21%, MAE at 16%, and MAPE at 11%. Mostly the public platforms were used for collecting data and IEEE, Springer and Science Direct were well-reputed journals for publishing articles for stock price predictions. Overall, this study provided useful insights into the current research methods and ideologies in analyzing market trends. Another study by (Subasi et al., 2021) used NASDAQ, NYSE, Nikkei, and FTSE indices and seven classifiers Random Forest, Bagging, AdaBoost, Decision Trees, SVM, K-NN, and ANN to understand the performance of supervised machine learning. The study showed that Random Forest and Bagging showed greater accuracy of 93%. They realized that by training on the leaked data, the accuracy was increasing. This study failed to provide uniformity in the use of technical indicators because each country had a specific indicator that played an important part in predicting stock trends. For example, FTSE: total market capitalization, NASDAQ: share weights and closing prices, NYSE: free-float market capitalization, and Nikkei: adjusted prices. This study revealed that stock prices are affected by economic supply and demands for a country. Subsequently, (Yelne and Theng, 2021) used both regression and classification methods of supervised machine learning using the Kaggle NSE50 dataset to forecast stock buy and sell signals. The results showed that the Regression method for decision trees, and logistic regression performed better than classification methods of the same. However, the study did not use any additional technical parameters and the size of the dataset was relatively small. Two methods were used i.e., LSTM and Regression with the former performing better. This study revealed that regression-based models are efficient for longer-term predictions where the size of the data frame is more than 5 lakhs. LSTM models help combat learning rate problems (i.e., vanishing gradient), caused due to large data by retaining the memory for the long term. Another study on LSTM by (Nelson et al., 2017) showed a Precision of 56% and an F1 score of 43% in stock price prediction using historical data. It stated that even after adding technical indicators and increasing the dimensions of the input data, there was no need to use dimensionality reduction methods for LSTM neural networks. (Qi, Khushi, and Poon, 2020) performed a study on the prediction of Forex prices. This study shed light on the use of the *Elliott wave*: for determining uptrend and downtrend and the Zig-*Zag indicator*: a signal for buying and selling that created sequences of events and importance on the use of LSTM and its variants like Bi-

LSTM and GRU. It also involved the use of other technical indicators for analysis. It showed that the GRU method for EUR/GBP currency had the best RMSE score of 1.5 ×10 $^{-3}$ and MAPE 0.12% for 15-minute interval data.

The stock market is a time series analysis. The linear model, ARIMA, and its variants AR, ARMA, and SARIMA are widely used for time series analysis. Usually, these models are less proficient in analyzing the underlying dynamics of the time-series data and do not perform well for multiple companies, i.e., the model works only for the company for which it was trained. (Hiransha et al., 2018) carried out a study on stock price prediction in global markets on companies under NSE and NYSE. The study revealed that non-linear deep learning models like CNN, MLP, RNN, and LSTM outperformed univariate time-series models like ARIMA. However, the study did not explore the use of the hybrid model. On the other hand, deep learning models are advantageous for financial analysis. DL models can be used on non-linear data and are good function approximators i.e., they can identify and learn the complex patterns of the dataset and provide generalized outputs. New test samples can be easily identified irrespective of being used in the training data. One single DL model can be used for predicting stock trends of a wide spectrum of datasets. This helps to achieve uniformity in applying models for various global markets. Another comparative study on predicting the next day closing price by (Vijh et al., 2020) depicts that ANN outperforms RF with an RMSE score of 0.42 and MAPE score of 0.77.

(Patel et al., 2015) conducted a study to compare the prediction performance of two Indian stocks i.e., Reliance, Infosys and two stock Indices i.e., S&P BSE Sensex and CNX Nifty using ANN, SVM, Random Forest, and Navies Bayes. They introduced Trend Deterministic Data Preparation Layer which converted continuous-valued inputs to discrete inputs ie -1 to +1 indicating probable future up or down movement based on inherent property. In the result, it was seen that all the models except ANN improved better with discrete data input (accuracy up to 90%). However, a clear understanding of why ANN had no effect on discrete input data was not provided.

(Nabipour et al., 2020) conducted a stock price prediction on the Tehran stock exchange using tree-based models Decision Tree, Bagging, and Random Forest. Adaboost, Gradient Boosting, XGBoost, and neural networks i.e., ANN, RNN, LSTM. The purpose of the study was to understand the differences in prediction outcomes for using input as continuous data and binary data. Deep learning models outperformed with an F1 score of 95%. This study shed light on the importance of data pre-processing for stock prediction. Binary data performed better than continuous data, also the running time for the binary data was much less. Deep learning models outperformed with an F score of 95% but this study does not have relevant kinds of literature in the given context for comparison. Reinforcement Learning (RL), an integration of Deep Neural Networks (DNN) with the art of human cognition decision-making ability is used for predicting stock prices. (Li, Ni, and Change, 2020) performed a study on RL. They used daily historical price and volume data for the US-based stocks and ETFs (Exchange-Trade Funds). The RL models used were a) Deep Q-Network (DQN), b) Double Deep Q-Network (DDQN), and c) Dueling Double Deep Q-Network (Dueling DDQN). The DQN model performed the best. This study revealed that these models have more intelligence than traditional models, they can easily adapt themselves to the volatile market environment and instantaneously respond to market fluctuations. However, there was inconsistency in the size of the dataset used in this research which compromised the results. Another study by (Meng and Khushi, 2019) showed that the RL model can be effectively used over baseline models to improve prediction accuracy and trade profitability. Conversely, there has been a paucity of literature in the comparative study of RL with ARIMA, LSTM, RF, NN, etc. (Iyyappan et al., 2022) introduced a system that predicted the stock prices in real-time using Algorithms i.e., recurrent neural network, Holt–Winters triple exponential implementation, and Recommendation system. The system was designed to get user information as the input which answered the Investment Amount, Duration, and Threshold spectrum of profit and loss the user can afford. The output machine would provide the closing price. This system, however, was not competent enough to provide low RMSE results, but it was successful in providing recommendations about stocks to buy or sell at a given interval.

2. B) Literature Review on Stock Predictions using Sentimental Analysis:

Since the stock market is a highly chaotic, volatile, and dynamic environment, it is quite challenging to get a certain framework for accurate predictions. A newer approach involves using human sentiments to better understand the market trends. The proceeding section discusses the works of researchers in relation to sentimental analysis.

Due to the rising impact of social media on our daily lives, a lot of focus and energy is given to sentiment analysis using Natural Language Processing (NLP), which is based on Twitter and news data. Sentiments i.e., the behavior of investors influence the market in the short-term causing a disconnection between the company's true value and stock price. (Jaggi et al., 2021) presented a study to understand the correlation between stock prices and the Stock Twits data. They proposed FinALBERT, i.e., an ALBERT-based model that labeled Stock Twits data for major FAANG companies. This study showed that there was a strong correlation between positive tweets and the price of the company increasing. (Nti, Adekoya and Weyori, 2020) used data from the Ghana Stock Exchange (GSE) and found a positive relation between the sentiments of people and the predictability of stock price movements. (Sharma et al., 2023) introduced a framework for analyzing the stock market using sentiments. The study used 15000 stock twits samples which were divided into positive, negative, and neutral data sets. The lexicon-based feature extraction method with the ABC algorithm was used to improve the model accuracy which turned up to 99.98% for positive data. However, this study did not take any statistical parameters of the financial market and was country-specific i.e., subject to change by political and economic situations of the country. (Malladi, 2022), study explains the practical use of supervised ML techniques to predict the US stock-market crashes at the time of COVID-19. The study used historical data from 1959-2020 and 134 economic variables from the FRED base. (Federal Reserve Economic Data). The study was successful in forecasting short-lived crashes, two months in advance.

However, sentimental analysis can be subjective to language bias leading to ambiguity, it might fail to detect comments of sarcasm and humor. Along with that, there can be data privacy and ethical considerations with respect to social media sources. The effects of sentiments on the market diminish over time and have no influence on long-term fluctuations. (Shah, Isah and Zulkernine, 2019) hence training on large-scale data can lead to discrepancy issues.

2. C) Literature Review on Stock Predictions using Technical Indicators:

Technical indicators analyze past prices and volume trends to forecast future prices. They play a significant role for traders and investors. They contribute to studying market trends, predicting future direction, determining entry, exit, and transition points, and helping to develop strategy and insights for market predictions. (Zhai, Hsu, and Halgamuge, 2007) pursued a study on using technical indicators and sentiments to predict the daily direction of BHP Billiton Ltd. (BHP.AX), an Australian stock. Four technical indicators were used Stochastic %K, Stochastic %D, Momentum, and Rate of Change. The study showed both profitability and SVM model performance increased. This study was done on only one stock company and did not compare the results before and after the use of technical indicators. In another similar study by (Orcharoen and Vateekul, 2018), they used an additional three indicators (i.e., William's %R, A/D Oscillator, and Disparity 5), and evaluated deep learning model performance i.e., CNN and LSTM. The use of indicators improved annualized returns based on trading simulations by 4.04%. (Kwon and Moon, 2007) validated the trading strategies based on technical analysis. They implemented a Neurogenetic Hybrid System on 36 stock companies in NYSE and NASDAQ with 13 years of historical data. The model showed better performance as compared to the "buy and hold" strategy. This study aimed to work on portfolio optimization and understand which stock to select and the number of shares in the stock to purchase or sell. This study used more than 20 features without performing PCA. (Neely et al., 2014) performed a comparative analysis of technical indicators and macroeconomic variables in predicting the Equity Risk Premium. Technical indicators had a better Sharpe ratio and significant predictive power as compared to the latter. This study used the Principal Component Analysis to identify important indicators and reduce the dimensionality. Overall, this study integrated the functioning of economic variables and their impact on technical indicators. Another study by (Ma and Yan, 2022) involved the use of 27 technical indicators to predict stock price changes using CNN in the Chinese stock market. The average forecasting accuracy of the CNN model was 69.89%. However, this study did not apply the algorithm to other datasets. Unlike the previous study, all the technical indicators were used for the analysis. PCA was not performed. Another research study by (Tanaka and Tokuoka, 2007) showed that adding ten technical indicators improved the prediction of eight stock tick data from NYSE to 82% in comparison to using a single indicator. Hence it can be seen from the literature that using technical indicators helps to improve the model performance.

2. D) Literature Review on Indian Stock Market Predictions and Literature Gap:

(Fathali, Kodia and Ben, 2022) performed a study on Indian Stock Market ie NIFTY-50 Index data. In the study, they used CNN, LSTM, and RNN to predict market trends. However, they used only the past 5 years of data and did not add any technical indicators. The results showed performance of LSTM was better than RNN and CNN for NIFTY-50 Index. Another study by (Sisodia et al., 2022) involved the use of Deep Learning based LSTM Algorithm. 10 years of data for 10 equities of NIFTY-50 Index was used for training and the highest accuracy was found to be 83.33% for SBI company. (Vikalp, Gupta, and Raj et al., 2018) conducted a study on the individual stock price prediction of NIFTY-50 Index. ANN with Back-propagation was used for modelling. However, they used only financial sector companies for their analysis. The study did not consider any other macro-economic variables. (Selvamuthu, Kumar and Mishra, 2019) used Neural network algorithms to compare stock price prediction between tick data and 15 minutes interval data. They found accuracy of tick data to be 99% as compared to 15 mins which accounted for 93%. This seemed to be an obvious finding as tick data involves more accurate datapoints. The algorithm was not implemented for other stock indices. (Raviraj et al., 2021) studied stock price prediction for the companies belonging to three sectors of NSE i.e., Power industry, IT industry and Pharma. Algorithms used were SimpleRNN, LSTM and GRU. The major drawback of the study was the high correlation between the training and testing data, i.e., standard deviation of 0.99, hence the results might not be precise for real world trading. (Banerjee and Mukherjee, 2022) conducted a study to predict the next 5 mins closing price using window sliding method, where they used non-parametric deep learning algorithms i.e., LSTM, GRU, BLSTM, BGRU, MLP for 3 companies listed in NSE. The average error predictions varied between 0.09% and 0.1%. However, they used only open, high, low, close, volume as input feature. They did not add any additional indicators.

To summarize, we reviewed literature works in financial research. There has been massive advancement in trading approach from traditional form of buying and selling to high frequency trading. The stock market is an interesting domain of research and has a wide spectrum of areas to explore for predictions. In the above section, we have tried to discuss literature papers with respect to deep learning and machine learning algorithms, sentimental analysis, technical indicators, and Indian stock market. All the research studies have been successful in predicting the accuracy of Machine learning and other Artificial Intelligence models, but there is a paucity of literature to understand the comparative analysis of algorithms in the Indian stock market for all the 50 companies under the NIFTY-50 Index. Previous research has performed analysis either on NIFTY-50 Index value or one sector listed in NSE or 4-5 equities in NIFTY-50. Also, technical indicators were not used for input data. Hence, in our research study, we try to use 10 algorithms for comparative analysis, more than 40 features for input data, and perform analysis on all the 50 equities listed in NIFTY-50 Index. This will help to get a wider perspective on the applicability of conventional machines and deep learning algorithms for trading purposes. In the next section, we describe the methodology used for the research study in detail.

Summary of Literature Review: In the (Table 3) below, we summarise all the papers reviewed in short.

References	Research Work	Insights
(Kumar, Sarangi and Verma, 2022)	Systematic literature review on stock predictions	ARIMA and NN were widely used algorithms
(Subasi et al., 2021)	Used 4 global indices and 7 classifiers	Random Forest and Bagging showed greater accuracy, dependent on macro-economic variables of country
(Yelne and Theng, 2021)	Compared regression and classification methods for supervised algorithms	Regression methods outperformed classification methods for same model
(Nelson et al., 2017)	Stock prediction using LSTM	Precision: 56% F1 Score: 43%
(Qi, Khushi, and Poon, 2020)	Forex price prediction using sequence-based events	GRU had lowest RMSE score for 15 min interval data for EUR/GBP currency
(Hiransha et al., 2018)	Comparative study between linear and non-linear models for NYSE and NSE	Non-linear time series model CNN, LSTM, RNN, MLP outperformed linear time series model ARIMA
(Vijh et al., 2020)	Close price prediction	ANN outperforms RF with RMSE 0.42
(Patel et al., 2015)	Trend deterministic data preparation layer to covert continuous input to discrete values -1 to +1	ANN improved better for discrete data
(Nabipour et al., 2020)	Stocking price prediction on Tehran stock exchange using tree-based and DL models and comparison between continuous input and binary input results	DL models with binary input data performed better with F1 score: 95%
(Li, Ni, and Change, 2020)	Stock prediction using different RL models	RL models adapted well as compared to DL models DQN model performed best
(Meng and Khushi, 2019)	Comparison of RL model with other conventional models	RL performes better and provides promising outcomes in terms of profitability
(Iyyapan et al., 2022)	Introduced system for real-time prediction	Results provided low RMSE, however not competent to use on versatile stock companies

(Jaggi et al., 2021)	Introduced FinALBERT to compare stock price with stock twits data for FAANG companies	Strong correlation was seen for stock tweets and stock price
(Nti, Adekoya and Weyori, 2020),	Correlation between sentiments and stock prices for GSE	Positive correlation
(Malladi, 2022)	Sentimental analysis and supervised ML to determine market crash	Successful in determining market crash due to COVID 19 with 2 months prediction window frame
(Zhai, Hsu, and Halgamuge, 2007)	Use of technical indicators along with sentiments for Australian stock exchange	Increase in SVM model performance with profitability
(Orcharoen and Vateekul, 2018)	Use of TI for DL models	Performance improved with use of TI
(Kwon and Moon, 2007)	Neurogenetic Hybrid System for portiolio optimization for NASDAQ and NYSE stocks	Adding >20 TI contributed for better results, however PCA was not applied
(Neely et al., 2014)	Comparative analysis of technical indicators and macroeconomic variables in predicting the Equity Risk Premium	TI had better Sharpe's ratio as compared to macro-economic variables
(Ma and Yan, 2022)	Stock price prediction in Chinese stock market using 27 TI and CNN	CNN accuracy 69.89%
(Tanaka and Tokuoka, 2007)	Use of 10 TI on tick data NYSE	Improvement in accuracy upto 82%
(Fathali, Kodia and Ben, 2022)	Used LSTM, CNN and RNN on NIFTY-50 Index (5 years of data without TI)	LSTM performed better than RNN and CNN
(Sisodia et al., 2022)	Deep Learning based LSTM Algorithm on 10 equity listed in NSE	SBI company accuracy was highest ie 83.33%
(Vikalp, Gupta, and Raj et al., 2018)	Performed analysis on financial sector of NIFTY-50 Index	Used ANN and back-prpogation, no use of macro-economic variables
(Selvamuthu, Kumar and Mishra, 2019)	Used Neural network algorithms to compare stock price prediction between tick data and 15 minutes interval data	Accuracy for: Tick data – 99% and 15 mins interval data – 93%
(Raviraj et al., 2021)	Implemented SimpleRNN, LSTM and GRU on 3 sectors of NSE	High correlation between training and test data ie SD: 0.99, results not precise
(Banerjee and Mukherjee, 2022)	Used LSTM, GRU, BLSTM, BGRU, MLP for 3 companies listed in NSE to predict next 5 mins closing price	The average error predictions varied between 0.09% and 0.1% No TI used in the study

Table 3: Literature Review Summary

Chapter 3: Methods

In the previous section, we reviewed multiple past papers and gained insights into the situation of financial research across the globe. There have been studies conducted on various machine learning and deep learning algorithms. In addition to that is the use of sentimental analysis for stock prediction. Studies have shown the importance of using technical indicators and their impact on model performance. The previous pieces of literature have effectively studied S&P500, and NASDAQ100 stock portfolios, however, we realized that there was a paucity of literature in performing comparative analysis for all 50 stock companies under the NIFTY-50 Index. Therefore, our research project aims to perform a comparative analysis of all 50 stock companies with machine learning and deep learning algorithms. In this section, we delve into a detailed explanation of our research approach and try to understand how the methodology aligns with the aim of this research study.

Research Aim: *Comparative analysis of various algorithmic models for the NIFTY-50 Index.*

This study is about short-term closing price prediction of stock using machine learning and deep learning algorithms. (Figure 3)

Given below are the methods that address the research objectives.

A) Dataset Collection B) Data Preparation C) Data Modelling and Optimization using machine learning and deep learning algorithms. D) Performance metrics

3. A) Data Collection:

The research study involves the analysis of fifty stock companies under the NIFTY-50 Index as stated in January 2023 on the National Stock Exchange India. These companies represent twelve sectors including Automobile, Banking, and IT that contribute to the Indian economy. (Girish, 2019) The dataset was collected from the open-source website Yahoo Finance (www.yahoofinance.com) (Table 4) In this study, we made use of Python 3 in the Google Colab environment for data analysis and computation. Historical data of the past 15 years i.e., from 1st August 2008 to 1st August 2023 was used for all fifty companies. The price in the dataset is expressed in INR Indian Rupee. The dataset was imported using the stock tickers from the Yahoo Finance library in Python. Each stock company under the NIFTY-50 Index was extracted as a sheet in the Excel file. Hence a single Excel file (.xlsx format) comprised of fifty sheets representing fifty stock companies. The data included day-wise trade history with Date, Open, High, Low, Close, Adj Close, and Volume. Further, a new column: ***Prev Close*** *(Previous close price)* was created by shifting the closing price by one day. *A list of 50 stock companies used in the study has been provided in the Appendix section.*

Dataset collection	Yahoo Finance
Computation	Python 3 in Google Colab
Data	50 Stock Tickers in NIFTY-50 INDEX
Time period	15 years data (01/08/2008-01/08/2023)
Data frame	1 Excel .xlsx file with 50 sheets for 50 stock companies
Libraries Used	Sklearn: Machine Learning Algorithms, Keras Tensor Flow: Deep Learning Algorithms Numpy and Pandas: Data Processing Matplotlib: Visualisation, TA-Lib: Technical Analysis

Table 4: Data Collection Summary & Libraries Used

Method Flow Chart:

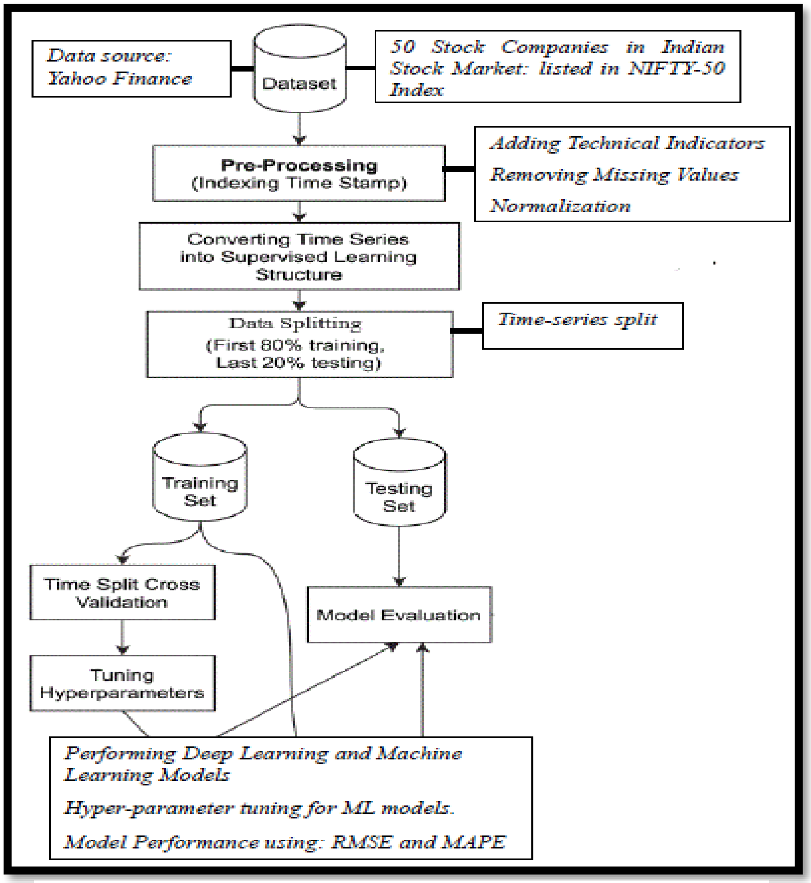

Data source: Yahoo Finance
Dataset
50 Stock Companies in Indian Stock Market: listed in NIFTY-50 Index
Pre-Processing (Indexing Time Stamp)
Adding Technical Indicators
Removing Missing Values
Normalization
Converting Time Series into Supervised Learning Structure
Data Splitting (First 80% training, Last 20% testing)
Time-series split
Training Set
Testing Set
Time Split Cross Validation
Tuning Hyperparameters
Model Evaluation
Performing Deep Learning and Machine Learning Models
Hyper-parameter tuning for ML models.
Model Performance using: RMSE and MAPE

Figure 3: Research Method Flow Chart

The research method type: Crisp- DM

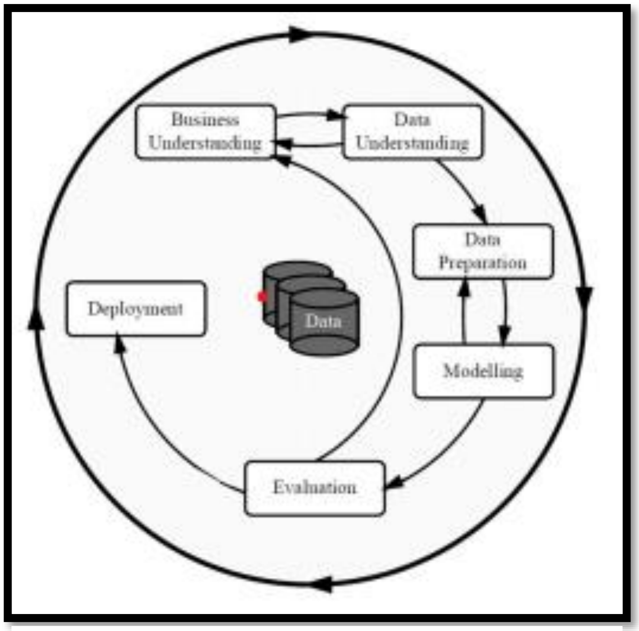


Figure 2: Cross Industry Process for Data Mining

(CS5704 Week 6 Teaching Materials, 2022)

In this research study, we made use of:

CRISP-DM Method: Involves formulating a research problem, applying data mining techniques and deploying a model to conclude final analysis. (Figure 2)

SMART goals: as an aid in formulating research methods. (Table 5) These goals include as follows: (CS5704 Week 3 Teaching Materials, 2022).

S: Specific	Definite Research Aim: Comparative analysis in Indian stock market
M: Measurable	Quantitative analysis Results evaluated using Performance Metrics i.e., RMSE and MAPE
A: Attainable and Achievable	Confirmed by previous research studies
R: Realistic and Resource	Secondary Dataset easily available on Yahoo Finance, Processing done in Python 3 Research Approach: Conclusive analysis
T: Time-bound	Project completed within time span of 3 months

Table 5: SMART Goals for Research Project Management

3. B) Data Preparation:

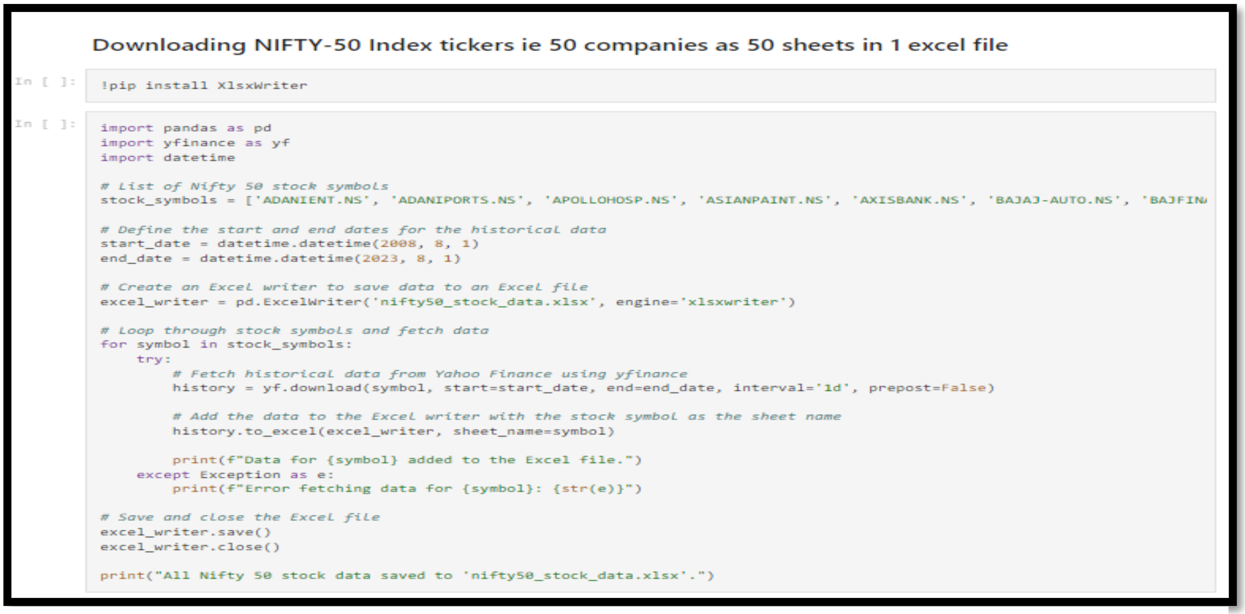
3.B). I Adding Previous Column:

A new column: ***Prev Close*** *(Previous close price)* was created by shifting the closing price of stocks in each row by one day.

3.B). II Adding Technical Indicators:

The technical indicators are mathematical formulas that are derived from the stock trading data. (Gao and Chai, 2018) Initially, each row in the dataset consisted of information from one day which is not sufficient to get insights about long-term trends. This can be improved by adding more features using the technical analysis (TA) library in Python. TA-Library helps to obtain new information from the past by creating newer variables. This enhances the quality of the original dataset. This study uses thirty-seven technical indicators using the TA-LIB package library. The thirty-seven indicators were derived from the original five indicators (ie open, low, high, close, volume) with distinct window sizes. The list of technical indicators used in the project is given in the (Table 6)

Stock market prediction is time-series data, non-stationary and volatile. The data complexity makes these data susceptible to noise. On the other hand, using technical indicators helps to smoothen the noise by giving trends and patterns in the data points. (Qi, Khushi, and Poon, 2020) However, there is a lack of sufficient literature to identify which technical indicators perform best for the NIFTY-50 Stock Index analysis. Hence, all the indicators are used as features to improve the model training. Each technical indicator has a specific Input window length i.e., a time frame Parameter used for deriving technical indicators There has been a study conducted by (Shynkevich et al., 2017) to understand the relation between the window time length for technical indicators and its effect on price trend predictions. It revealed that the model has the highest performance measures when the window length is equal to the forecasting horizon. Since data used in this study is from the past 15 years, technical indicators are calculated at different window lengths of 5,10, 50, 100, and 200. (Figure 4)



Downloading NIFTY-50 Index tickers ie 50 companies as 50 sheets in 1 excel file

```
In [ ]: !pip install XlsxWriter
```

```
In [ ]: import pandas as pd
        import yfinance as yf
        import datetime

        # List of Nifty 50 stock symbols
        stock_symbols = ['ADANIENT.NS', 'ADANIPORTS.NS', 'APOLLOHOSP.NS', 'ASIANPAINT.NS', 'AXISBANK.NS', 'BAJAJ-AUTO.NS', 'BAJFIN

        # Define the start and end dates for the historical data
        start_date = datetime.datetime(2008, 8, 1)
        end_date = datetime.datetime(2023, 8, 1)

        # Create an Excel writer to save data to an Excel file
        excel_writer = pd.ExcelWriter('nifty50_stock_data.xlsx', engine='xlsxwriter')

        # Loop through stock symbols and fetch data
        for symbol in stock_symbols:
            try:
                # Fetch historical data from Yahoo Finance using yfinance
                history = yf.download(symbol, start=start_date, end=end_date, interval='1d', prepost=False)

                # Add the data to the Excel writer with the stock symbol as the sheet name
                history.to_excel(excel_writer, sheet_name=symbol)

                print(f"Data for {symbol} added to the Excel file.")
            except Exception as e:
                print(f"Error fetching data for {symbol}: {str(e)}")

        # Save and close the Excel file
        excel_writer.save()
        excel_writer.close()

        print("All Nifty 50 stock data saved to 'nifty50_stock_data.xlsx'.")
```

Figure 4: Data Preparation

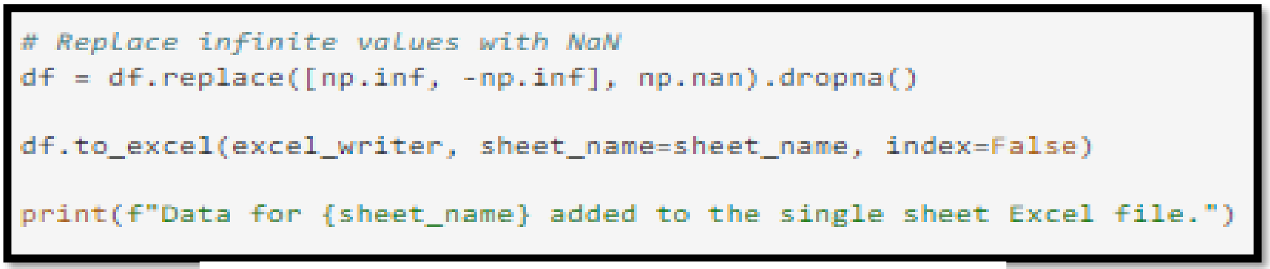
Technical Indicators	Description	Formula
Simple Moving Average (SMA) ***Windowlength:*5,10,20,50,100,200**	The average price over a specific period	*SMA= (A1+A2+…+An)/n* *An: price of an asset at period* *n: number of total periods*
Exponential Moving Average (EMA) **Window length: 5,10,20**	Weighted moving average that measures a trend for recent data	*EMA=Price(t)×k+EMA(y)×(1−k)* *t=today, y=yesterday, N=number of days in EMA, k=2÷(N+1)*
Moving Average Convergence/ Divergence (MACD)	Trend and cross-over indicator that is calculated by the difference in EMA on 12th day and 26th day	*MACD=12-Period EMA − 26-Period EMA*
Relative Strength Index (RSI)	Momentum oscillator that measures the speed and change of price movements	*RSI = 100 – [100 / (1 + (Average of Upward Price Change / Average of Downward Price Change))]* *RSI value oscillates between 0-100*
Vortex Indicator (VI)	VI is composed of two lines - an uptrend line (VI+) and a downtrend line (VI-) and is used to spot trend reversals and confirm current trends.	*Upward movement (VM+):* *current high - previous low* *Downward movement (VM-): current low - previous high*
Bollinger Bands	Divide into upper, middle, and lower bands as per the standard deviation, helps to understand the volatility	*Upper band = 20-day SMA + (20-day SD x 2)* *Middle band = 20-day SMA* *Lower band = 20-day SMA – (20-day SD x 2)*
Average True Range	Gives insights into the degree of price volatility	*(Previous ATR(n−1) +TR)/n* *Where n = number of periods* *TR=True range*
Stochastic Oscillator %*K*	Compares closing price to previous price n days prior and informs if stock is overbought or oversold	%*K*=100×*CP*−*L*14/*H*14−*L*14 *CP*=Most recent closing price *L*14=Lowest price of the 14 previous trading sessions *H*14=Highest price of the same 14 previous trading sessions
Williams %	Depicts the relationship between the current closing price and the high and low prices over the previous n days usually 14 days, indicates overbought and oversold	Wiliams %R = (Highest High−Lowest Low / Highest High−Close) where: Highest High=Highest price in the lookback period, typically 14 days. Close = most recent closing price. Lowest Low=Lowest price in the lookback period, typically 14 days.
Price rate-of-change (ROC)	Relative difference between the closing price on the day of forecast and the closing price n days previously	ROC=(*Closing* Price *p /* Closing Price *p*−*n*)×100 Where Closing Price*p:* Closing price of most recent period Closing Price*p*−*n* = Closing price *n* periods beforemost recent period

Table 6: Technical Indicators (Shynkevich, 2017)

3.B). III Removing Missing Values:

Data cleaning ensures that the integrity of the stock data is well maintained. This reduces the chances of skewness and biases in the data. Overall, data quality assures better model performance. (Zhao and Zhao, 2021) carried out a research study where they used four methods for feature selection method on a dataset with 1500 stock companies, 49 input features, and 9 years of data. These methods were a) deleting rows with missing values b) removing unique values c) removing values of high correlation and low importance. In a study on short-term prediction using Machine learning conducted by (Huang, Capretz and Ho, 2021), quarterly data of 100 companies under the S&P 100 Index for the past 22 years was used. Technical indicators were also used, which accounted for up to large blocks i.e., 50% of missing values. They removed all the rows with missing values. In real-world trading, the multi-variate financial data is unstable with the presence of multiple noise and missing numbers. Another study by (Boonpeng and Jeatrakul, 2014) explained two methods for maintaining the integrity of data quality prior to ANN i.e., first by removing the missing values or mean-based imputation and second by normalizing the data.

In the given research study, many technical indicators had values from day 50, day 100, day 200, etc. Hence, there were multiple rows that had NAN value. These rows containing NAN were identified and removed from the dataset to prevent data discrepancy issues while modeling. The dataset was complete and there were no other missing values present. (Figure 6)

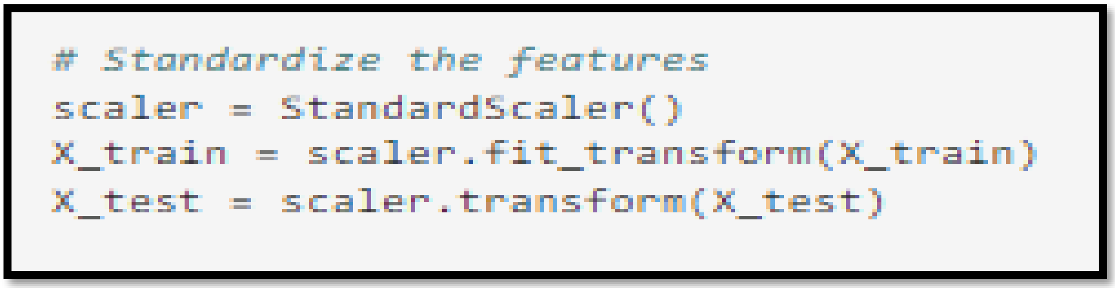


```
# Replace infinite values with NaN
df = df.replace([np.inf, -np.inf], np.nan).dropna()

df.to_excel(excel_writer, sheet_name=sheet_name, index=False)

print(f"Data for {sheet_name} added to the single sheet Excel file.")
```

Figure 6 Missing Values



```
# Standardize the features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Figure 5 Normalisation - Min-Max Scaling

3.B). IV Normalization:

The dataset used for the study involved data from fifty different stock companies. Each stock company had a distinct range of input features. This variation in the scale of input features may cause bias and disparity in the modelling results. Machine learning and Deep learning models function better by feature normalization. There are higher chances of finer coefficients, optimized networks, adjusted model weights, and improved outcomes by training data. (Gómez, Martínez and Martín, 2022) stated algorithms that use gradient descent of hyper-parameter optimization need data scaling. The difference in the scale of the input features causes changes in the step sizes in the calculation of gradient descent which leads to an increase in computational power. The algorithms perform well if they develop a meaningful relationship between the target variable and explanatory variables. Therefore, data scaling is a crucial step in data pre-processing. Data scaling is mainly done by Standardization and Normalization. This study uses Normalization i.e., the process of rescaling the numerical features in the dataset in such a way that the new values fit in a range of 0 and 1. (Alam et al., 2020) This can be achieved by using a Min-Max Scaler, Z-Score, Decimal Scaling, and Log transformation. We used Min-Max Scaler for data normalization. (Figure 5) It is a widely used technique in financial predictions. It helps to bring uniformity to the stock data. (Hiransha et al., 2018) Studies have shown a reduction in MAPE using the Min-Max Scaling method. (Boonpeng and Jeatrakul, 2014) The algorithms perform well if they develop a meaningful relationship between the target variable and explanatory variables.

Normalization: x norm = (x − x min) / (x max − x min)
where x norm is the normalized value
x min and x max are the minimum and maximum value in the training dataset.

3.B). V Prediction Model:

In this study, we used five deep-learning models and five machine-learning models to predict the 10th-day Closing price of fifty stocks under the NIFTY-50 Index as given in the table. Since the stock market is highly volatile and influenced by political and sentimental situations, it was not preferable to predict long-term duration prices for multiple companies. Hence only short-term prediction (10 days) was studied.

All the coding was executed on all 50 companies using *for-loop* iterating through *50 sheets in 1 Excel file.*

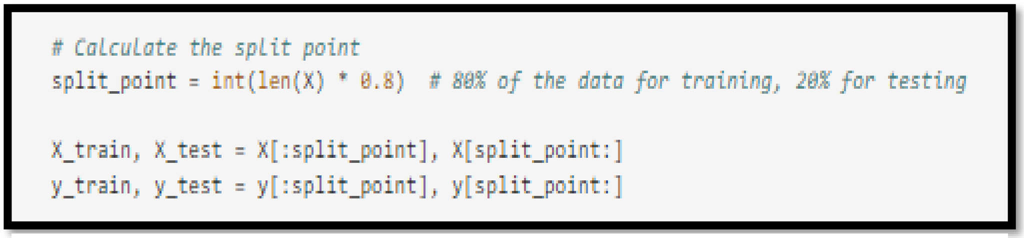
Input features: The dataset consisting of open, high, low, adj close, volume, and previous close along with 37 indicators is used as the input data. (Figure 8)

Output: The 10th-day closing price is used for prediction.

Training and Testing data:

The entire dataset is divided into 80% training data and 20% testing data as per the chronological sequence. (Figure 7)

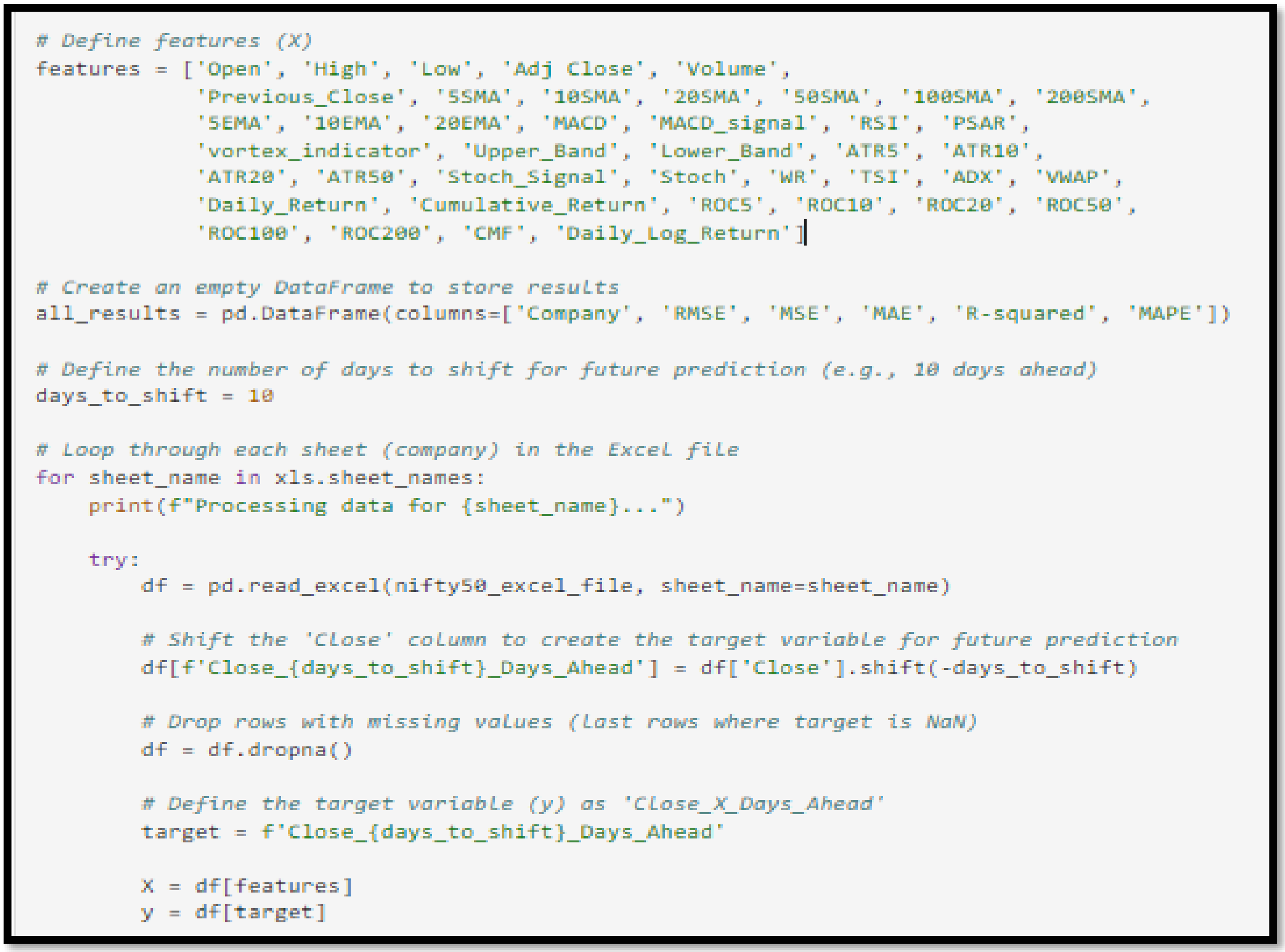
We cannot use K-fold cross-validation for randomly sampling the data into train and test because stock prediction is a time-series data. Due to temporal dependence and auto-correlation, the data points are not independent of each other. Hence methods like time-series splitting and rolling window cross-validation are usually used. This complies with future forecasting and prevents the problem of overfitting.



```
# Calculate the split point
split_point = int(len(X) * 0.8)  # 80% of the data for training, 20% for testing

X_train, X_test = X[:split_point], X[split_point:]
y_train, y_test = y[:split_point], y[split_point:]
```

Figure 7: Training and Testing Data



```
# Define features (X)
features = ['Open', 'High', 'Low', 'Adj Close', 'Volume',
            'Previous_Close', '5SMA', '10SMA', '20SMA', '50SMA', '100SMA', '200SMA',
            '5EMA', '10EMA', '20EMA', 'MACD', 'MACD_signal', 'RSI', 'PSAR',
            'vortex_indicator', 'Upper_Band', 'Lower_Band', 'ATR5', 'ATR10',
            'ATR20', 'ATR50', 'Stoch_Signal', 'Stoch', 'WR', 'TSI', 'ADX', 'VWAP',
            'Daily_Return', 'Cumulative_Return', 'ROC5', 'ROC10', 'ROC20', 'ROC50',
            'ROC100', 'ROC200', 'CMF', 'Daily_Log_Return']

# Create an empty DataFrame to store results
all_results = pd.DataFrame(columns=['Company', 'RMSE', 'MSE', 'MAE', 'R-squared', 'MAPE'])

# Define the number of days to shift for future prediction (e.g., 10 days ahead)
days_to_shift = 10

# Loop through each sheet (company) in the Excel file
for sheet_name in xls.sheet_names:
    print(f"Processing data for {sheet_name}...")

    try:
        df = pd.read_excel(nifty50_excel_file, sheet_name=sheet_name)

        # Shift the 'Close' column to create the target variable for future prediction
        df[f'Close_{days_to_shift}_Days_Ahead'] = df['Close'].shift(-days_to_shift)

        # Drop rows with missing values (last rows where target is NaN)
        df = df.dropna()

        # Define the target variable (y) as 'Close_X_Days_Ahead'
        target = f'Close_{days_to_shift}_Days_Ahead'

        X = df[features]
        y = df[target]
```

Figure 8: Input and Target Variable

These were the algorithms used in the research study:

Deep Learning Algorithms	**Machine Learning Algorithms**
Artificial Neural Network (ANN)	Decision Tree (DT)
Convolutional Neural Network (CNN)	Support Vector Machine (SVM)
Hybrid Model (CNN-ANN)	Random Forest (RF)
Long-Short-Term- Memory (LSTM)	Gradient Boost
Multi-Layer Perceptron (MLP)	XG Boost

Table 7: Data Modelling Algorithms

3 C) Data Modelling:

Given below are the deep learning and machine learning algorithms used in the study. (Table 7)

3 C). I ANN:

Artificial Neural Networks (ANN) are extensively used for forecasting and approximation functions. They have single or multi-layer neural networks. There is an input layer, an output layer, and two hidden layers. Nodes are the interconnection between each layer. (Nabipour et al., 2020) ANN can approximate a large class of functions with relatively higher degrees of accuracy. This is an additional advantage over other non-linear models. ()

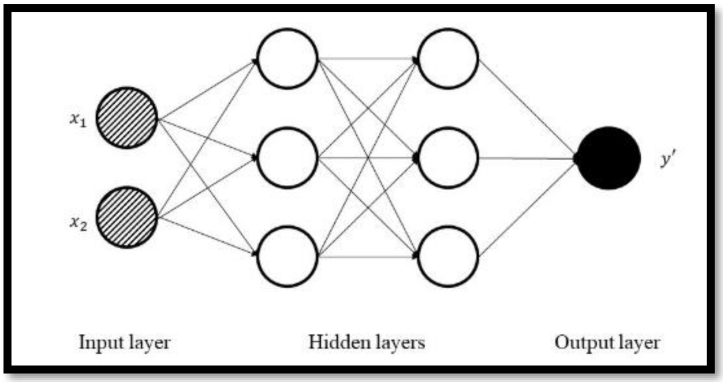


Figure 9: Artificial Neural Networks (Nabipour et al., 2020)

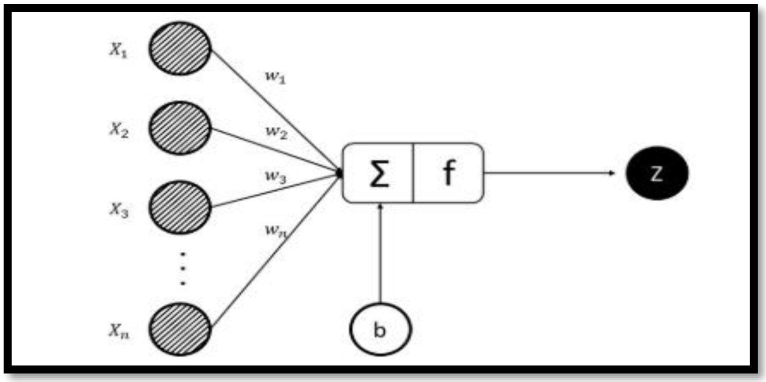


Figure 10: Input and Output in ANN (Nabipour et al., 2020)

$$Z = \mathrm{f}\,(\mathrm{x.w} + \mathrm{b}) = \mathrm{f}\left(\sum_{i=1}^{n} x_i^T w_i + b\right).$$

The above equation denotes the relationship between nodes, bias, and weights. The total weighted sum of input layer is passed to another node in the next layer through the non-linear activation function.

Where, X1, X2, X3, …., Xn = Inputs

w1, w2, w3 ,…, wn = Weights

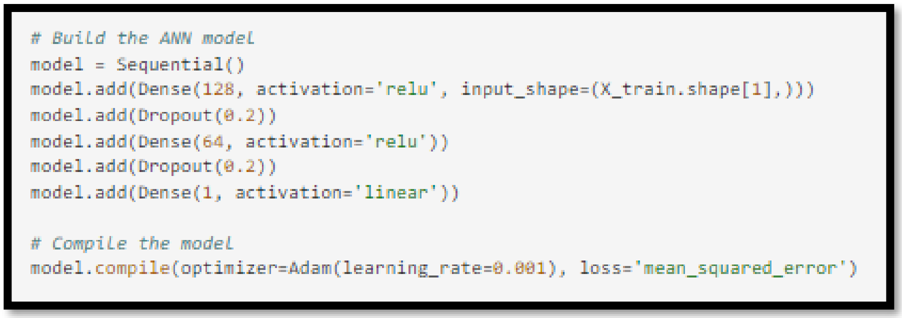
n = final node: number of inputs

f = activation function

z = output

Library	Keras Library (Tensor Flow)
Number of hidden layers and neurons	128 neurons in the first hidden layer, 64 neurons in the second hidden layer
Activation function	ReLU (Rectified Linear Unit) in both the layers
Drop-out rate	0.2% (To prevent overfitting)
Learning rate (Adam optimizer)	0.001

Table 8: ANN Parameters



```
# Build the ANN model
model = Sequential()
model.add(Dense(128, activation='relu', input_shape=(X_train.shape[1],)))
model.add(Dropout(0.2))
model.add(Dense(64, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(1, activation='linear'))

# Compile the model
model.compile(optimizer=Adam(learning_rate=0.001), loss='mean_squared_error')
```

3 C). II CNN:

Convolutional Neural Networks (CNN) are a class of deep neural networks consisting of three layers, a convolutional layer, a pooling layer, and a fully connected layer. They are widely used in computer vision. The input layer in CNN takes the raw data. The second layer i.e., the Convolutional layer uses kernels to detect complex patterns in the data. This is followed by activation using (ReLU Rectified Linear Unit), responsible for adding non-linearity. The pooling layer contributes to reducing computation power by reducing the dimensions of data. Dense layers form interconnected convolutions which are then flattened to 1-D vector. Lastly, the output layer produces the result. (Figure 11)

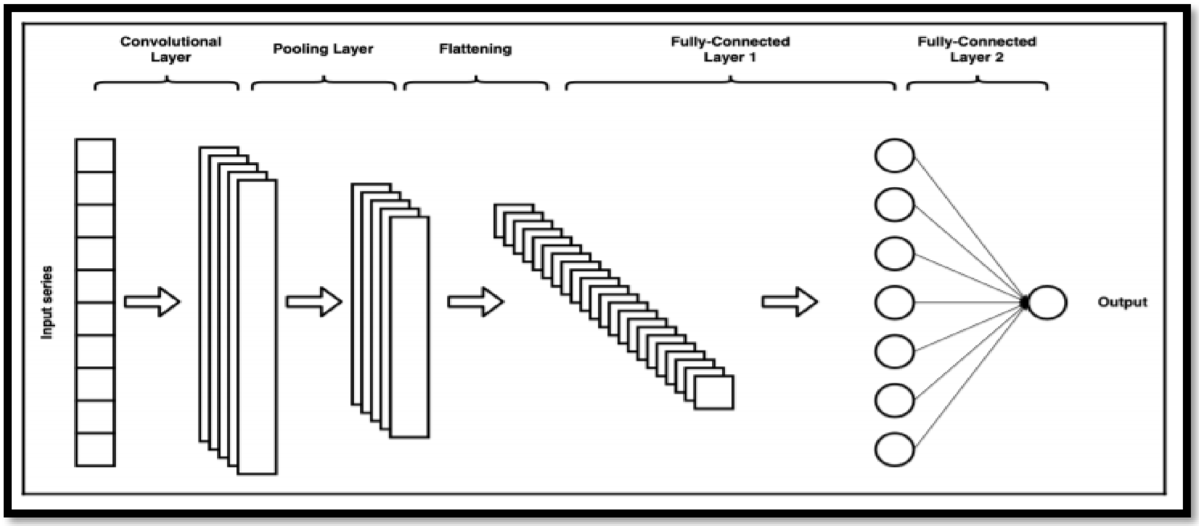


Figure 11: CNN

Table 9: CNN Parameters

Parameter	Description
Input layer	Receives the input data
1st Convolutional Layer	Filters: 32, Kernel Size: 3, Activation: ReLU
1st Max Pooling Layer	Pool size: 2 (For reducing dimensions)
2nd Convolutional Layer	Filters: 64, Kernel Size: 3, Activation: ReLU
2nd Max Pooling Layer	Pool size: 2
Flatten layer	Absent
1st Dense Layer	Neurons: 128, Activation: ReLU
1st Dropout Layer	Dropout rate: 20% (For Regularization)
2nd Dense Layer	Filters: 64, Activation: ReLU
2nd Dropout Layer	Dropout rate: 20% (For Regularization)
Dense Layer: Output	Neurons: 1, Activation: Linear (Regression type)

3 C). III MLP:

Multi-Layer Perceptron (MLP) is a feed-forward Neural Network where the information moves in the forward direction. Neurons are connected in a set architecture which is in the form of layers. The input layer and output layers are connected by hidden layers. (Figure 12) The neuron in the given layer receives output from the previous layer and passes it forwards. The neurons are interlinked through weighted matrix. (Banerjee and Mukherjee, 2022) Since it a fundamentally like ANN, MLP's are universal function approximators. With the help of non-linear activation functions, it can detect non-linear data patterns. However, in respect to hyper-parameter tuning, MLP are prone to vanishing gradients.

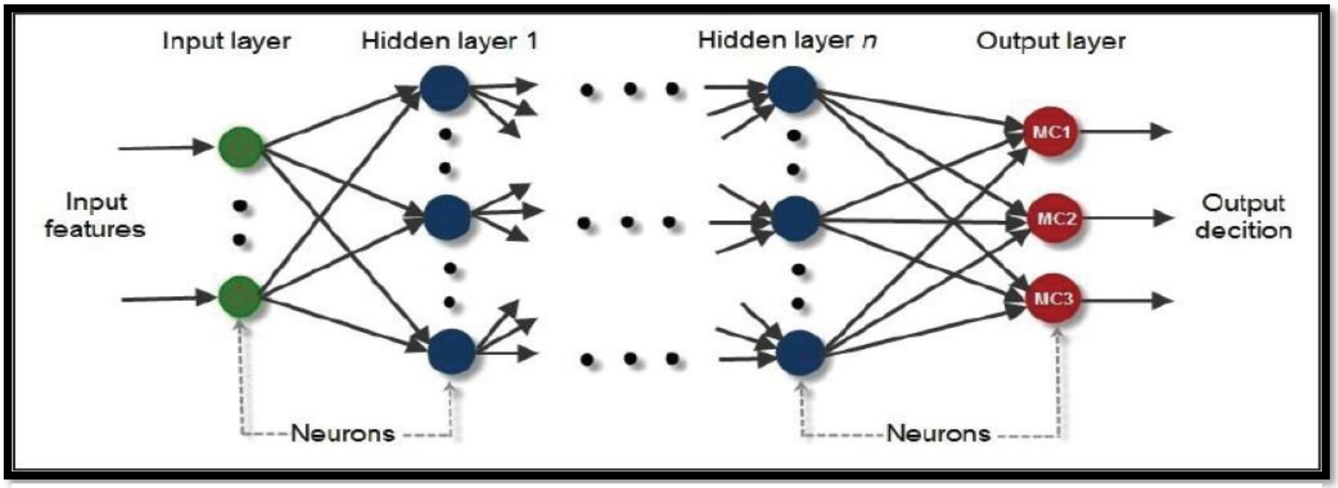


Figure 12: MLP (Banerjee and Mukherjee, 2022)

Table 10: MLP Parameters

Parameter	Description
Hidden layer sizes	First layer: 128 Second layer: 64
Activation	Rectified Linear Unit (ReLU) activation function
Solver	Adam optimizer
Maximum number of Iterations	1000
Validation fraction	20% of data
Early stopping	True
Number of iter change for early stopping	10

3 C). IV LSTM:

Long Short-Term Memory is a type of RNN, the only difference is that the neuron in LSTM has memory. It is widely applicable in stock prediction time-series analysis, image recognition, speech, and voice classification due to its ability to store and discard information with the help of gates. LSTM neurons have three gates namely, input gate: is responsible for adding information, forget gate: deletes information which is not required by the model and output gate: selects information and displays as output. (Figure 13)

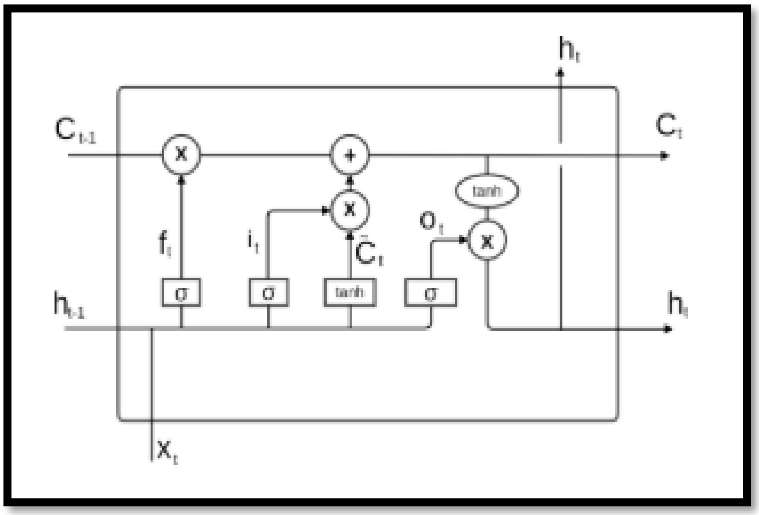


Figure 13: LSTM Cell (Hiransha et al., 2018)

Ct−1, Ct : conveyor belt that carries information from the previous cell to the present one.

f_t : forget gate layer = uses sigmoid activation to decide which information to store in the cell

h_t : Output = sigmoid gate (Ot) multiplied by tanh,

i_t : input gate= combines sigmoid and tanh

(C·): new updated value of cell formed by addition of sigmoid and tanh

LSTM Layers	Parameters
LSTM Layer 1	Units: 128 Activation Function: ReLU Return Sequences: True Dropout Rate: 0.2
LSTM Layer 2	Units: 64 Activation Function: ReLU Return Sequences: True Dropout Rate: 0.2
LSTM Layer 3	Units: 32 Activation Function: ReLU Return Sequences: False Dropout Rate: 0.2
Output Dense Layer	Units: 1 Activation Function: Linear

Table 11: LSTM Parameters

3 C). V Machine Learning Models:

1. ***Support Vector Machine (SVM):***
SVM Regression is considered a significant technique for non-linear stock prediction in time-series analysis. (Grigoryan, 2016) SVM rapidly adapts to high-dimensional feature space with the use of kernel functions. The important purpose of SVM is to identify the maximum margin hyperplane (Patel et al., 2015) Advantages of using the SVM model are: a) Scalability- can be easily applied on large datasets with the help of Kernel Approximators b) Using RBF kernel can help to detect hidden patterns in the non-linear data c) SVM is robust to outliers and reduces error d) Prevents overfitting by use of hyper-parameters.

2. ***Decision Tree (DT):***
DT are non-parametric, supervised machine learning method. It is a representation of a tree which splits as per the feature significance. The tree structure is formed by *Root:* initial feature that divides data, *Leaf node:* Terminal node, makes numerical predictions for regression problems, *Decision nodes:* the intermediate between root and leaf node, *splitting criterion:* depends on performance error i.e., MSE in regression. (Figure 15)

3. ***Random Forest (RF):***
The process of training multiple decision tree together i.e., ensemble learning. The predictive performances are calculated by taking average of subsets of decision trees. RF reduces variances present in individual decision trees, improves the error score by using larger number of bootstrapped samples, prevents overfitting. (Figure 14)

4. ***Gradient Boost (GB):***
This technique is like RF i.e., ensemble of weak learners. The difference is in the prediction i.e., gradient of error: The new learner adapts to the residual error by previous learner and improves the decision gradient wise.

5. **XG Boost (XG):**
XG Boost (Extreme Gradient Boost) is an updated technique of GB. This has more processing speed, better efficiency, optimization, and regularization techniques as compared to traditional GB. (Figure 16)

Machine Learning Models	**Parameters used**
Support Vector Machine: Regressor	Hyper-parameter Tuning: Grid Search CV Regularization Parameter Value: 0.1, 1, 10 Epsilon Value: 0.01, 0.1, 1 Kernel: RBF (Radical Basis Function) Gamma: scale, auto, 0.1, 1
Decision Tree	Max depth: None, 15, 30, 45 Min_sample_split: 4, 10, 20 Min_sample_leaf: 2, 4, 8 max_features: 'auto', 'sqrt', 'log2'

	criterion: 'mse', 'mae'
Random Forest	n_estimators: 200, max_depth: 30 min_samples_split: 2 min_samples_leaf: 1 max_features: 'sqrt'

XG Boost	n_estimators=300
	max_depth=10
	learning_rate=0.1
	subsample=1.0
	colsample_bytree=1.0

Gradient Boost	n_estimators=800
	learning_rate=0.01
	max_depth=12

Table 12: Machine Learning Parameters using in Data Modelling

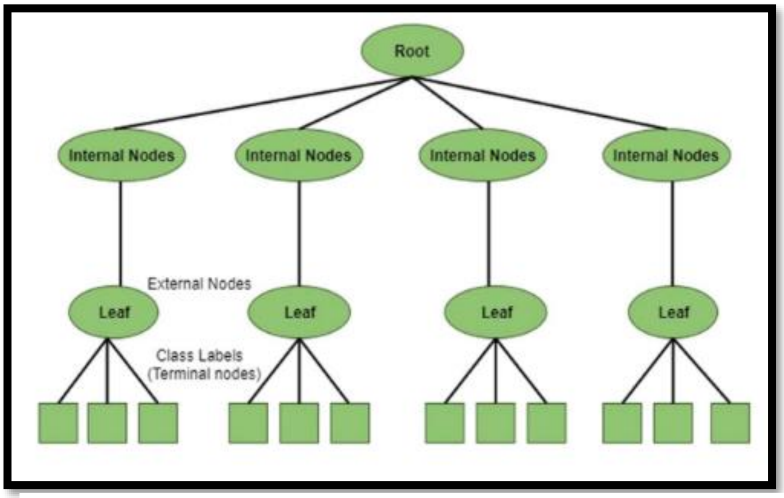


Figure 15: Decision Tree Structure (Alam et al., 2020)

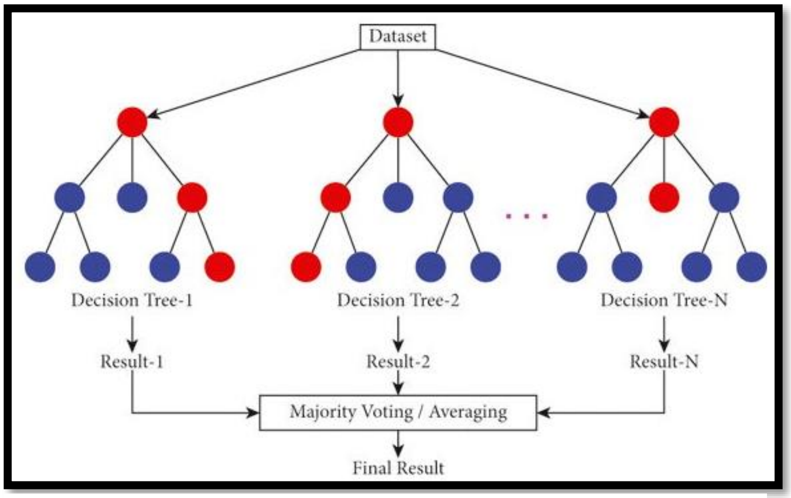


Figure 14: Random Forest

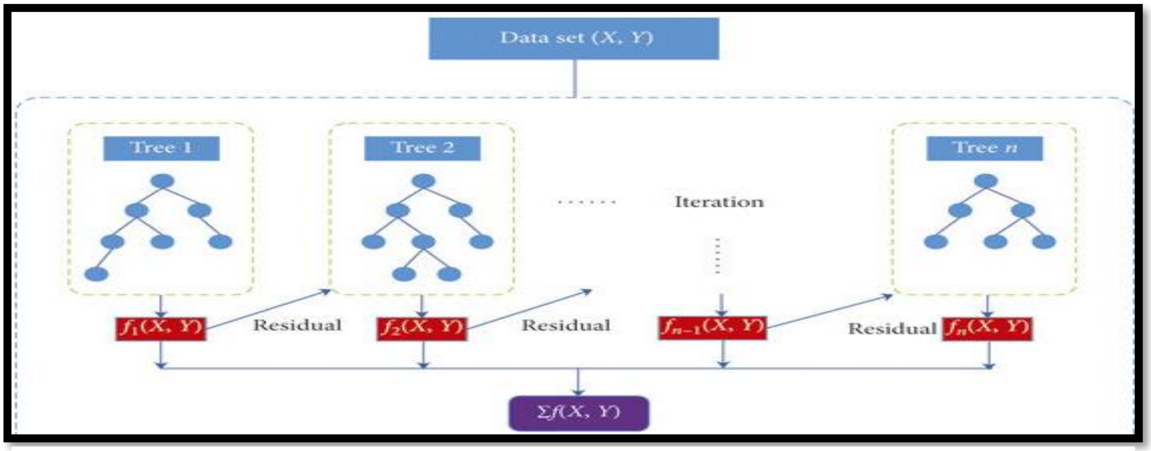


Figure 16: XG Boost (Han, 2019)

3 D) Performance Metrics:

3 D). I Root Mean Square Error (RMSE):

RMSE is a widely used performance metric for predicting accuracy in the Regression type of studies. (Jain, Gupta, and Moghe, 2018), (Althelaya et al., 2018), (Narayana et al., 2022), (Rajab and Sharma, 2015) It is defined as the standard deviation from prediction errors in Regression. (Nabipour et al., 2020). The distance between actual values and predicted values is calculated as residuals or prediction errors. Mean square error (MSE) is the square of prediction errors divided by mean. RMSE is the square root of MSE. (Figure 17) RMSE is computationally easy to use especially for larger datasets. However, it is susceptible to outliers and may not give accurate results for large differences in the magnitude of data. Hence scaling data prior to modelling is crucial. The lower the value of RMSE, the better the fit of the model.

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

Figure 17: RMSE Formula

Where ŷi = Predicted Value, yi = Actual Value, n= total number of data points, i= time

3 D). II Mean Absolute Percentage Error (MAPE):

MAPE is calculated by taking the average difference between the actual value and the predicted value and dividing it by the actual value. This value is taken as an average and converted to a percentage. (Figure 18) MAPE is used as a performance metric in time-series analysis. (Hu, Zhao, and Khushi, 2021), (Hoque and Aljamaan, 2021), (Hiransha et al., 2018) The conversion to percentage is a better representation of data for the investors. MAPE considers taking relative error metric which makes it suitable to identify underlying patterns in the time-series data. It can be easily compared across multiple algorithms hence the lower the value of MAPE, the better the model performance.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \,.\, 100\%$$

Figure 18: MAPE Formula

Where ŷi = Predicted Value, yi = Actual Value, n= total number of data points, i= time

3 E) Project Planning and Risk Management:

A project is a well-defined task that has a definite starting point and an ending point. Project management is the process to ensure that the end goal is achieved efficiently. (CS5704 Week 6 Teaching Materials, 2022) Planning is a crucial step in project management. It includes recognizing the relevant project deliverables and providing specifications on the execution of project tasks for meeting the end goal and objectives. There are many methods available for planning a project. These include Top-Down planning, Bottom-up approach, Waterfall method etc. (CS5704 Week 6 Teaching Materials, 2022) We implemented the Waterfall method i.e., linear method initially which was followed by iterative changes and flexibility with research journey. Secondly, the achievable milestones in the research process were determined with the help of GNATT chart. This includes eight major stages from reading literature papers, writing introduction, literature reviews, finding gaps and proposing research methods, ethical approval, followed by analyzing results, using hyper-parameters, writing discussions, seeking supervisor guidance, implementing essential changes, concluding the study. Also, throughout dissertation managing risk is extremely crucial. Given below is the method for managing Project risks. (Table 13) (CS5704 Week 6 Teaching Materials, 2022)

3*3 Matrix with Scores 3=High risk, 2= Medium risk, 1= Low risk

Risk ID	**Risk Description**	**Probability**	**Impact**	**Priority**	**Mitigation Plan**
1	Computational difficulties: working with big data	3	3	3	Keep the version updated and use GPU cluster for better data processing. Data back up Grey box software testing
2	High Budget	1	2	2	Regular monitoring of project finances
3	Delayed Time	3	3	3	Time Management Skills
4	Health Issues	1	2	3	Disciplined life with good mental health can help relieve stress and

Table 13: Project Risk Management Matrix

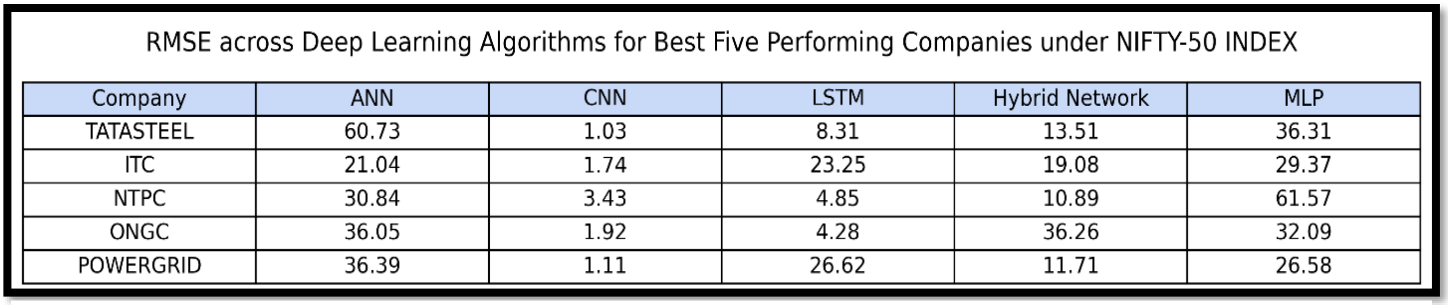
Chapter 4: Results

We will discuss the findings of the research study in the following section. This research study aims to analyze the closing price of the stock companies under the NIFTY-50 Index. There have been 10 models used in the project out of which 5 are machine learning models and 5 are deep learning models. A total of 42 variables are taken as input data and the 10th day closing price is the target variable. This study uses the ***Regression*** type of Supervised Learning where the target variable is the Numerical datatype.

The algorithms were processed using *for loop* iterating through all the sheets in the single Excel file. Each sheet represented data of the single stock company. A total of 50 stock companies under the NIFTY-50 Index (NSEI) were taken into consideration. Each algorithm for individual companies was compared with respect to the RMSE and MAPE scores. The tables below depict the results of the best 5 companies with the lowest values of RMSE and MAPE scores across all the machine learning and deep learning algorithms. It was seen that Tata Steel company which is in the metal sector performed best across all the algorithms with RMSE values ranging from 1.03 to 68.11 and MAPE values ranging from 1.9 to 57.1. The second best-performing company is Power Grid which is in the energy-power sector, with RMSE value ranging from 1.11 to 86.54 and MAPE value ranging from 1.45 to 35.28. Other companies that performed relatively well were ITC *(Consumer Goods sector)*, NTPC *(Energy Power sector),* and ONGC *(Oil and Natural Gas cooperation in Energy-Oil and Gas sector)*

4 A) Deep Learning Algorithms:

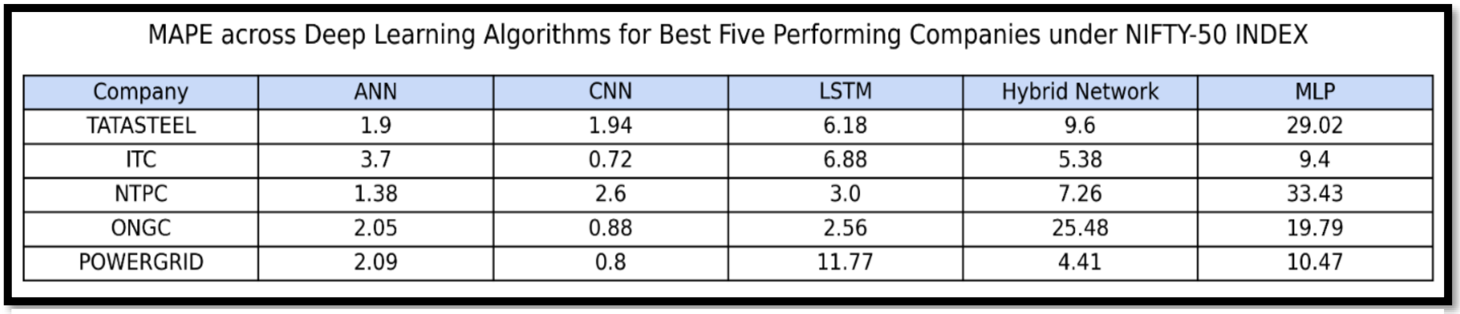
A total of 5 algorithms were implemented for 50 stock companies i.e., Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), Hybrid Network of CNN-ANN, Multi-layer Perceptron (MLP). The performance of ***CNN*** was the best for all the companies in comparison to other algorithms. The RMSE score for CNN ranged from 1.03 to 3.43. The lowest RMSE in the entire data analysis was for ***Tata Steel company: 1.03***. (Table 14) However, other deep-learning models like ANN, MLP, and LSTM did not perform as well as CNN. Even the Hybrid Network i.e., a combination of CNN with ANN had an RMSE score ranging from 11.71 to 36.26. With respect to MAPE, CNN and ANN models showed relatively similar performances. (Table 15) The lowest MAPE score was for ITC, ONGC, and PowerGrid which were 0.72,0.88, and 0.80 respectively. Multi-Layer Perceptron (MLP) did not perform well with high RMSE and MAPE values up to 61.57. (Figure 19, 20)



RMSE across Deep Learning Algorithms for Best Five Performing Companies under NIFTY-50 INDEX

Company	ANN	CNN	LSTM	Hybrid Network	MLP
TATASTEEL	60.73	1.03	8.31	13.51	36.31
ITC	21.04	1.74	23.25	19.08	29.37
NTPC	30.84	3.43	4.85	10.89	61.57
ONGC	36.05	1.92	4.28	36.26	32.09
POWERGRID	36.39	1.11	26.62	11.71	26.58

Table 14: RMSE score - Deep Learning Algorithms



MAPE across Deep Learning Algorithms for Best Five Performing Companies under NIFTY-50 INDEX

Company	ANN	CNN	LSTM	Hybrid Network	MLP
TATASTEEL	1.9	1.94	6.18	9.6	29.02
ITC	3.7	0.72	6.88	5.38	9.4
NTPC	1.38	2.6	3.0	7.26	33.43
ONGC	2.05	0.88	2.56	25.48	19.79
POWERGRID	2.09	0.8	11.77	4.41	10.47

Table 15: MAPE score - Deep Learning Algorithms

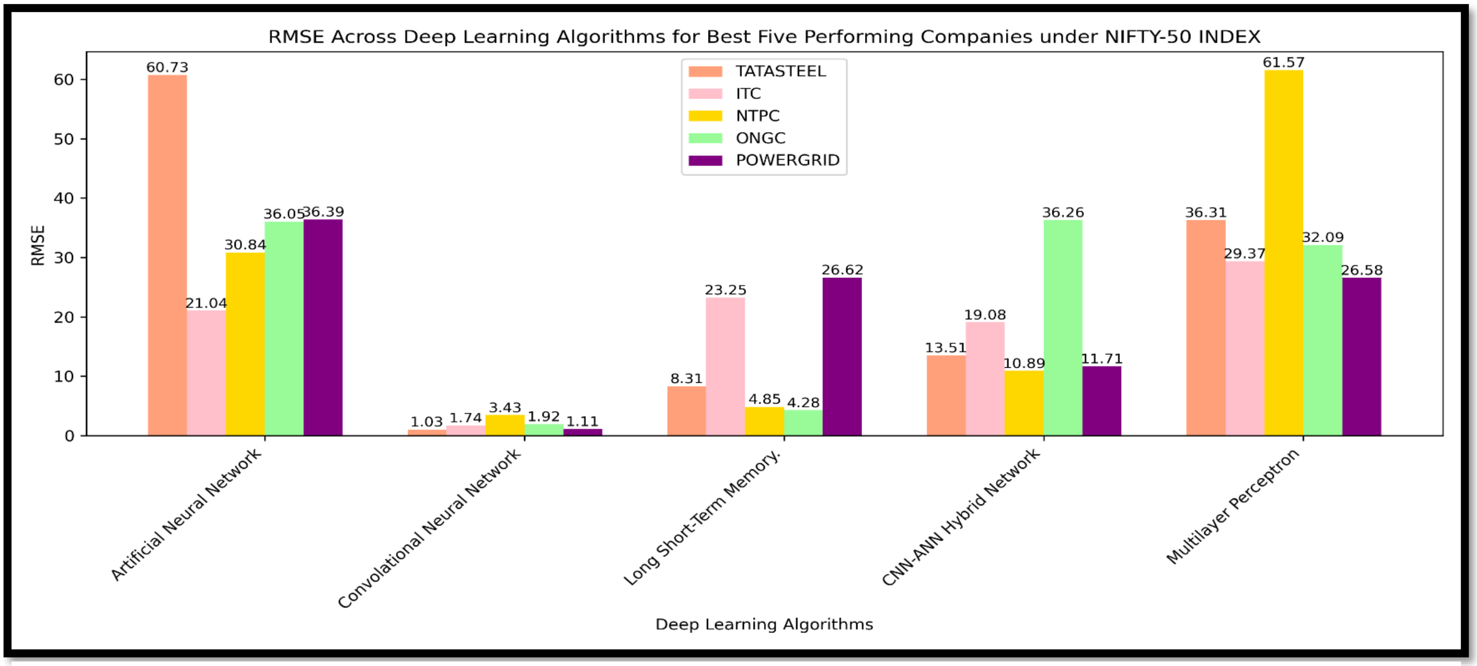


Figure 19: RMSE vs Deep Learning Algorithms:

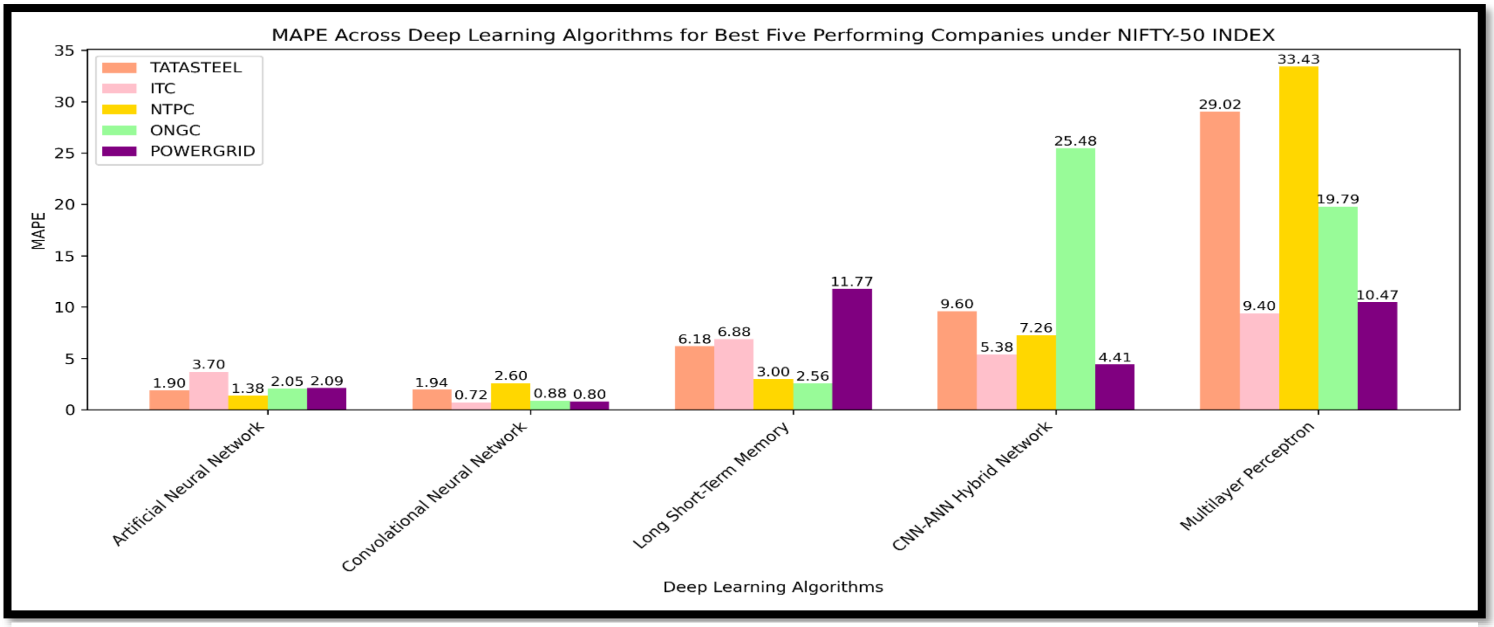


Figure 20: MAPE v/s Deep Learning Algorithms

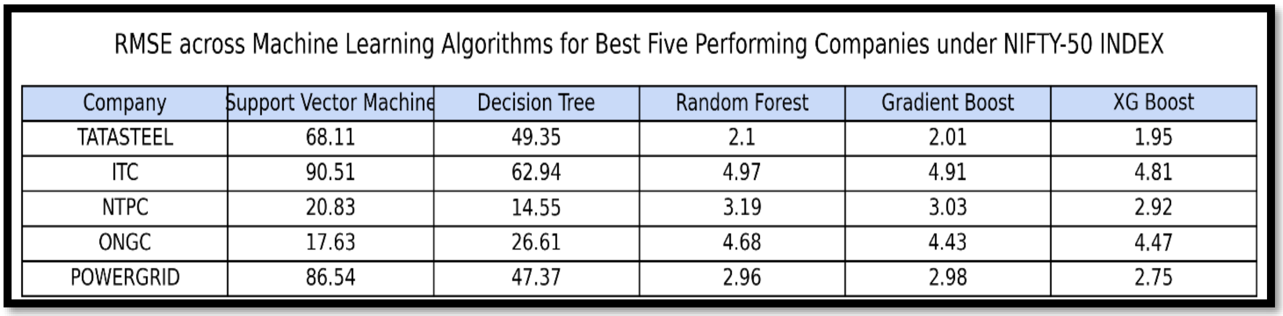
4 B) Machine Learning Algorithms:

In the given research study, 5 supervised ML algorithms were used which were Support Vector Machine, Decision Trees, Random Forest, Gradient Boost, and XG Boost. The lowest RMSE score was for Tata Steel company using the XG Boost Algorithm i.e., 1.95 and the lowest MAPE score was for PowerGrid company ie 1.45. Random Forest XG. (Table 16, 17)

However, the performance results for Random Forest, Gradient Boost, and XG Boost were quite overlapping. Decision Tree and Support Vector Machine results with optimized hyper-parameters derived from Grid Search CV were not equivalent to the rest of the models. SVM had a relatively higher value of RMSE scores accounting up to 86.54 for Power Grid company and 90.51 for ITC company. MAPE values were lower for Random Forest, Gradient Boost, and XG Boost ie between 1.45-2.65, and were higher for SVM and Decision trees ie 57.1 and 39.97 respectively.

Summary:

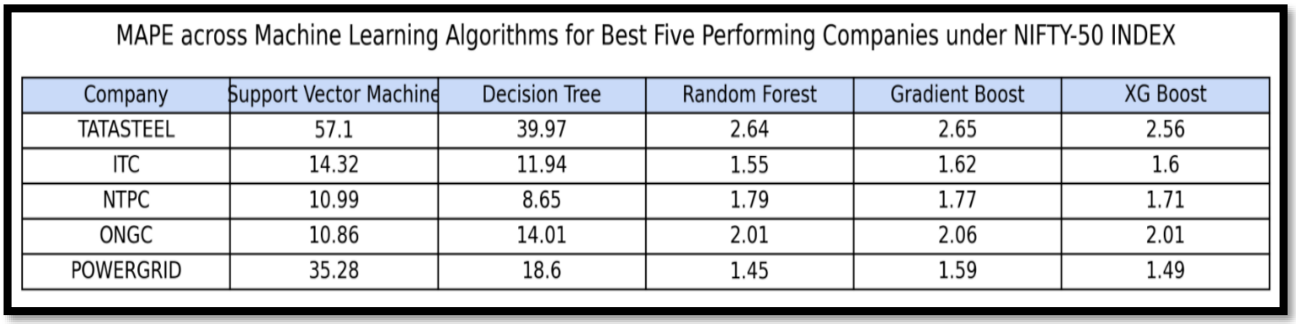
	Deep Learning	***Machine Learning***
Best Models:	ANN > CNN > LSTM > Hybrid > MLP	XG Boost > Gradient Boost > Random Forest > Decision Tree > Support Vector Machine
Best performing Companies:	TATA Steel followed by NTPC, ONGC, PowerGrid ITC	PowerGrid and TATA Steel followed by NTPC, ONGC, and ITC



RMSE across Machine Learning Algorithms for Best Five Performing Companies under NIFTY-50 INDEX

Company	Support Vector Machine	Decision Tree	Random Forest	Gradient Boost	XG Boost
TATASTEEL	68.11	49.35	2.1	2.01	1.95
ITC	90.51	62.94	4.97	4.91	4.81
NTPC	20.83	14.55	3.19	3.03	2.92
ONGC	17.63	26.61	4.68	4.43	4.47
POWERGRID	86.54	47.37	2.96	2.98	2.75

Table 16: RMSE score - Machine Learning Algorithms



MAPE across Machine Learning Algorithms for Best Five Performing Companies under NIFTY-50 INDEX

Company	Support Vector Machine	Decision Tree	Random Forest	Gradient Boost	XG Boost
TATASTEEL	57.1	39.97	2.64	2.65	2.56
ITC	14.32	11.94	1.55	1.62	1.6
NTPC	10.99	8.65	1.79	1.77	1.71
ONGC	10.86	14.01	2.01	2.06	2.01
POWERGRID	35.28	18.6	1.45	1.59	1.49

Table 17: MAPE score - Machine Learning Algorithms

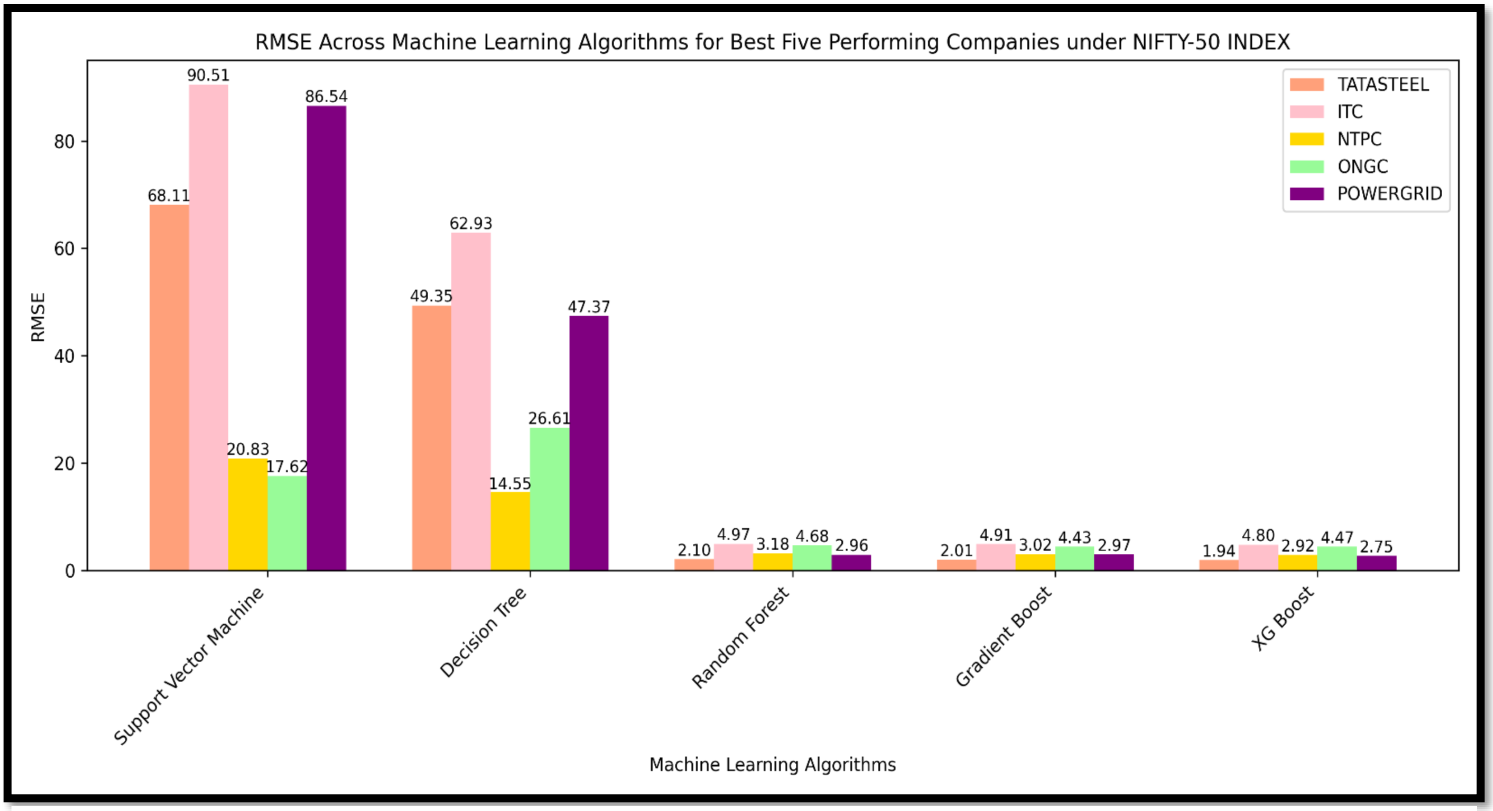


Figure 21: RMSE v/s Machine Leaning Algorithms

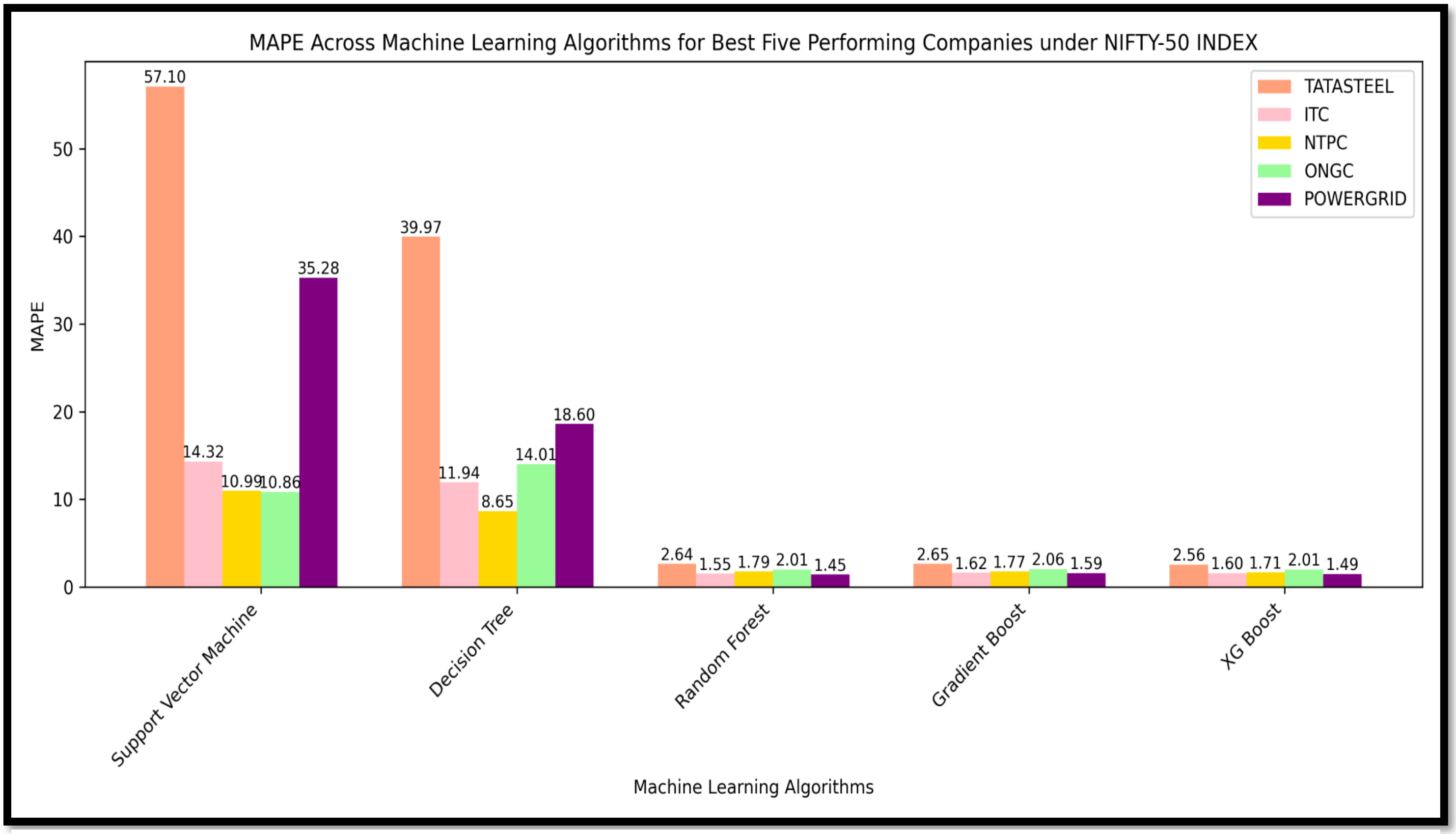


Figure 22: MAPE v/s Machine Learning Algorithms

4 C) Comparing Actual v/s Predicted Results:

For all the stock companies, the training is done on the first 80% of the data, and testing is done on the subsequent 20% of data i.e., chronologically since it is a time-series data. The actual closing price for the 10th-day v/s predicted closing price for the 10th day is plotted below using different algorithms for TATA STEEL stock company. The pink line represents the Actual 10th-day Closing Price, and the green line represents the Predicted 10th-day Closing price. It can be seen from the graphs below that the Performance of CNN (figure 23) is better than ANN (figure 22). MLP (figure 25) and LSTM (figure 24) have more differences in actual and predicted prices which can be confirmed by high RMSE and MAPE scores.

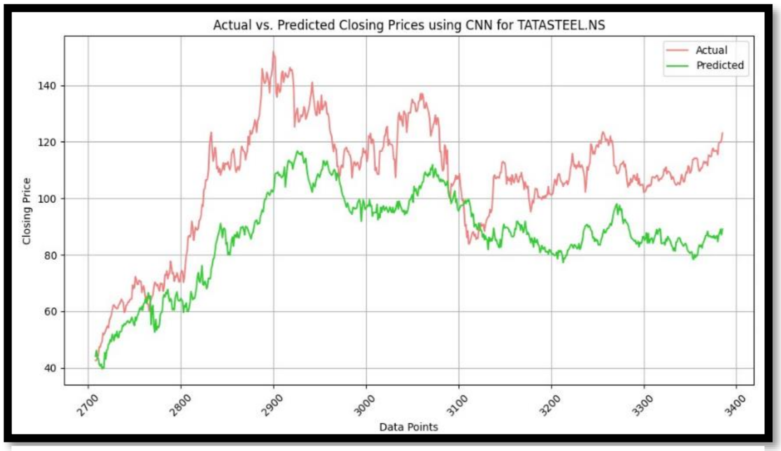


Figure 21: CNN- Actual v/s Predicted Closing Price

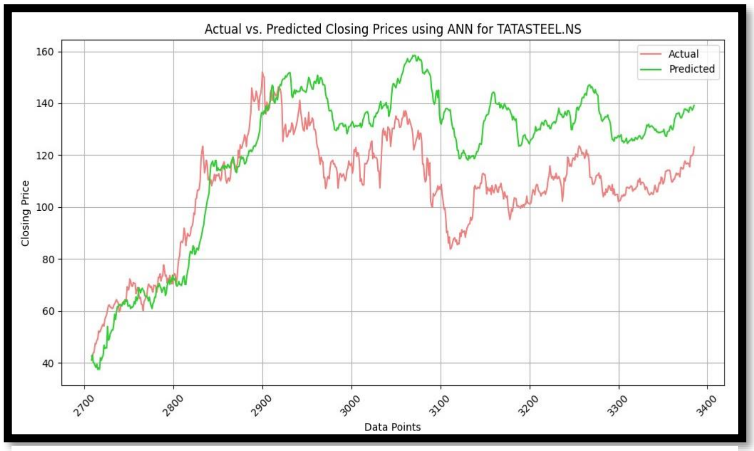


Figure 20: ANN - Actual v/s Predicted Closing Price

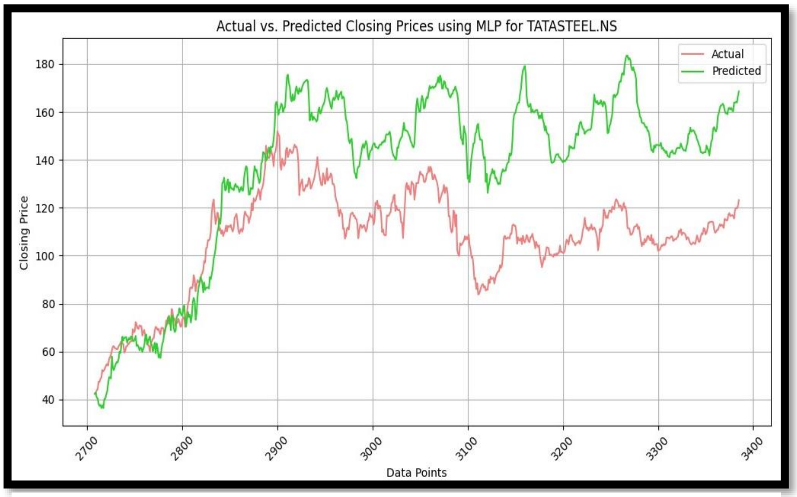


Figure 23: MLP- Actual v/s Predicted Closing Price

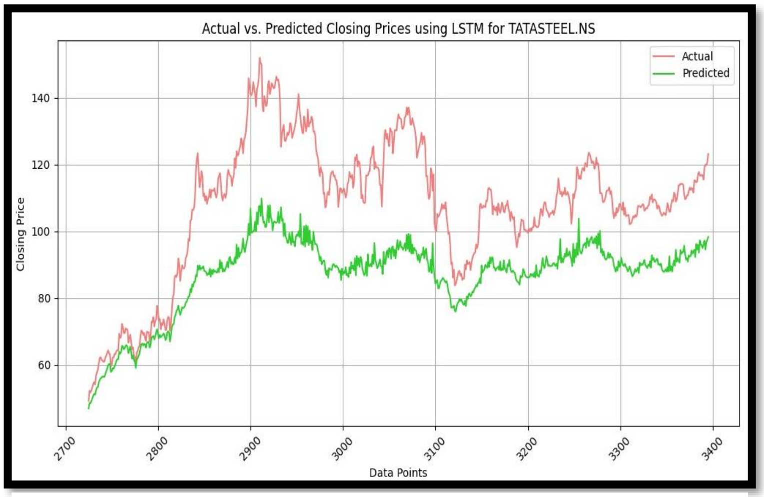


Figure 22: LSTM- Actual v/s Predicted Closing Price

4 D) Feature Importance:

A total of 42 features were taken as input data. It was seen that each algorithmic model had different features that played a significant role. For example, it can be seen from the graphs below that Adjacent Close was the most important feature for SVM with RBF Kernel but least important for Random Forest. (Figure 26,27) However, overall Simple Moving Averages at various window lengths and MACD signals played a vital role in almost all the models. It can be seen from (figure 28) that 25% of technical indicators were correlated to each other. These technical indicators were Simple Moving Averages, Exponential Moving Average, ATR, and Bollinger Bands had an approximate correlation of 1.

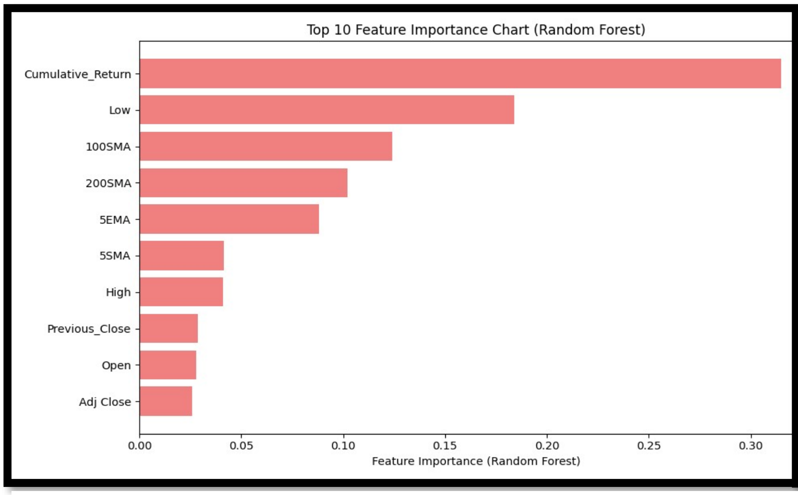


Figure 24: Feature Importance- RF

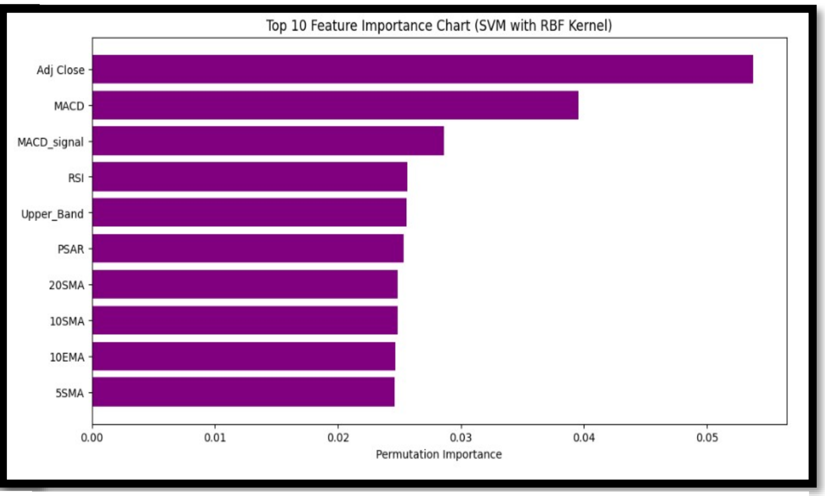


Figure 25: Feature Importance- SVM

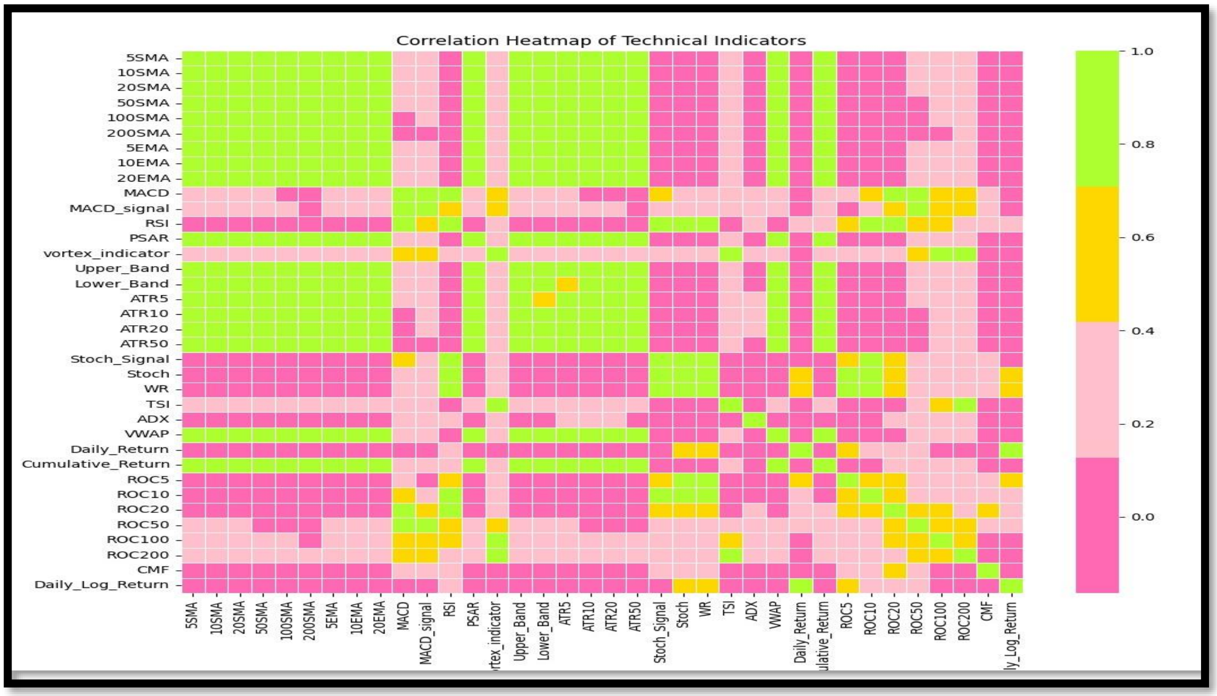


Figure 26: Heat Map- Technical Indicators

Chapter 5: Discussion

The given section discusses the results of the research study. It focuses on comparative analysis and model predictions. It starts by explaining the results of the project followed by providing an overview of challenges associated with stock price prediction in the Indian stock market. Later, this section ends with a short summary of insights derived from the analysis.

Stock market data is an example of non-stationary data. At a given time, the market can show various trends, cycles, random walks, or a combination of all three. (Patel et al., 2015) Due to the volatile nature of the stock market, traders and analysts are more curious about short-term predictions (i.e., weekly, and monthly predictions) as compared to longer-term predictions. Predicting the stock market is a complex task. In our research study, we aimed to perform a comparative analysis of all the 50 stock companies under the NIFTY-50 Index. We predicted short-term closing prices using a total of ten deep-learning and machine-learning algorithms. The results were compared using two performance metrics i.e., RMSE and MAPE. The results were diversely ranged for all 50 companies. With a wide spectrum of RMSE and MAPE scores for different companies, we sorted the best 5 companies that responded well to the algorithms. These companies were PowerGrid, TATA Steel, ITC, NTPC and ONGC. The rest of the companies had relatively higher scores for performance metrics. The past 15 years of historical data along with 36 technical indicators was taken as input data. The correlation heat map for technical indicators revealed that approximately 25% of indicators had a correlation of 1. These indicators were SMA, EMA, ATR, and Bollinger Bands. Since these indicators show a high correlation, i.e., less variance in the dataset, they could have been eliminated from the input feature prior to modelling. This could prevent the problem of multi-collinearity. On the other hand, it was seen from the feature importance graph that Cumulative return, SMA, and low price played an important role for Random Forest and Adjacent close, MACD, RSI, and Bollinger Band were important indicators for SVM with Grid Search CV, and RBF Kernel. This is contradicted by the results of the correlation i.e., the indicators that showed high collinearity turned out to be important features for many algorithms. With respect to the algorithms used in the study, it was seen that the overall performance of CNN, XG Boost, and Random Forest was the best with lower values of RMSE scores. Usually, CNN is widely used for image analysis which involves spatial data. Conventionally, time-series models like ARIMA, and LSTM (RNN) perform better for time-series forecasting due to their temporal and sequential nature. SVM is especially for non-linear data due to its hyper-plane marginal separation. However, SVM with optimized hyper-parameters had higher scores of RMSE and MAPE, which contradicts the fact that SVM is the best model. Alternatively, RF and XG Boost had relatively better prediction performance with lower error scores. This accounts for the fact that RF and XG Boost use Ensemble methods which combine multiple decision trees i.e., creating a strong learner. Secondly, stock market data may contain diverse prices due to its volatile nature. Ensemble methods can quickly adapt to outliers and prevent their impact on predictions. Additionally, ensemble methods have built-in regularization mechanisms that prevent the problem of over-fitting and biases. The parallel computation in these algorithms helps in faster processing time especially when the dataset size is large. The computational time to process 50 stock companies in 1 Excel file for RF and XG Boost was less than 10 mins whereas for LSTM and CNN it was more than 6 hours. LSTM and MLP had high RMSE scores. MLP failed to develop sequentially temporal patterns between the high-dimension input features. Also, MLPs do not have large memories like LSTM thereby they do not connect to the long-term dependencies and hidden patterns in the data. LSTM works well for large datasets; however, it did not perform well on diverse stock companies. The comparative analysis of all companies under the NIFTY-50 Index showed diversified results for different algorithms. Despite normalizing the values and performing data pre-processing techniques, it is a challenging task to bring uniformity to the results. The stock market is a massive domain of the financial market with greater extensibility. Previous works of literature have analyzed their proposed algorithms for less than 5 stock companies. This gives promising assurance that the algorithms work well. But when it comes to modelling for 50 stock companies, there is a broad spectrum of variations in the results. Each stock company is distinct and responds uniquely to multiple factors which include the company's fundamental values, political influence, technical indicators, the economic condition of the country, etc. One algorithm that works well for a country might not perform well for another country.

Algorithmic trading systems (ATS) have created a shift in the functionality of the stock market system. In today's era of Industry 4.0, using a computational platform for trading is inevitable. This system has multiple benefits that include low cost of operation, reduced latency, high-frequency trading, diversification, scalability, a 24*7 trading platform, risk management, back-testing optimization, less dependence on human emotions, and a more logical approach. At the same time, predictions by ATS influence panic selling and disturb the decorum of the market. As a result, the market overreacts to this and in turn brings newer results in the predictions. This ever-repeating cycle goes on. However, retail investors face challenges in adopting algorithmic trading because of its technical nature and technical built-in system requirements. With flourishing in research studies, newer algorithms are regularly introduced in the market. There have been thousands of articles published globally on stock price predictions including the stock market of multiple countries. However, there is no success when it comes to the comparison of the reliability and validity of these algorithms and prediction techniques. It has been observed that the algorithms which can generate huge profits are not shared publicly. These algorithms are kept confidential and private. It accounts for the self-defeating nature of the stock market. Hence, the research study and strategy behind such algorithms are generally never published. (Shah, Isah and Zulkernine, 2019) Similarly, with respect to sentimental analysis, the data produced on social media can be fake and unreliable. Dependencies on fake data can cause reverberation in the market by negative reinforcement. Alternatively, using the annual or quarterly reports issued by the companies can help to comprehend the company's future trends. (e.g., 10-Q and 10-K) (Shah, Isah, and Zulkernine, 2019)

To summarize, this research study on comparative analysis of 50 stock companies in the Indian stock market provided three major insights, a) Each company is formed by amalgamation of various factors which include fundamental, country's economic and political factors. Finding relevant explanatory variables as input features is unique for each stock company. b) The stock company data is highly non-linear, diverse, dynamic, volatile, subjected to sentiments, and unstructured. Fitting this data to one perfect algorithm might not work on a long-term basis. This market is subject to change with time, hence relying on the best algorithm for short-term predictions is relatively profitable as compared to long-term predictions. c) Grossly, this research study was successful in finding the best 5 companies that performed well for all the machine learning and deep learning algorithms. Further analysis with back-testing can give more clarity on the real-time trading for these companies.

Chapter 6: Conclusion

This section summarizes the main points of the research study on Indian stock price prediction, followed by key insights and learnings from this study, research limitations and future scope of studies that evolves from the dissertation.

6 A) Summary of the Dissertation:

The main purpose of the study was to perform a comparative analysis between 50 Stock Companies under NIFTY-50 Index i.e., Indian stock market using various deep learning and machine learning algorithms. 50 Stock companies in NIFTY-50 Index as per current year 2022 were taken as datasets. Data was collected using Python script from Yahoo Finance for the last 15 years. Each dataset was prepared as a single sheet in one Excel file representing single stock company and all the sheets were processed simultaneously through iterations in .xlsx file using for-loop. 37 Technical Indicators were added to the existing data frames for all companies. Later, data was cleaned by removing missing values. The input data consisted of a total of 43 features and the target variable to be predicted was the 10th day closing price. Since large datasets with multiple companies were used, there was variation in the scale of data, hence data scaling using Normalization i.e., Min-Max Scaler was done. This was followed by splitting the data into 80% training and 20% testing chronologically. Deep learning models: ANN, CNN, LSTM, MLP, Hybrid Network CNN-ANN and Machine learning models: DT, RF, XG Boost, Gradient Boost and SVM were used. The models were evaluated using two Performance Metrics measures i.e., RMSE and MAPE. Out of 50 Indian stock companies, best five performing companies for all the algorithms were TATA Steel, ITC, NTPC, PowerGrid, ONGC. Out of these companies, three companies belong to the Energy and Power supply sector. The best algorithms with lowest RMSE scores i.e., in range of (1.4 - 4.6) were CNN, XG Boost, Gradient Boost and RF. However, the Hybrid Model, LSTM, MLP did not perform well with high RMSE scores i.e., up to 70. The Machine learning models were optimized using Grid Search CV, but this did not improve the performance error score for SVM using RBF kernel. The actual vs. predicted graph for test datasets were plotted for different algorithms. CNN and ANN had better predictions as compared to MLP and LSTM. It was seen that each Algorithm had varied important features. Overall, this study performed an analysis of all 50 stock companies.

6 B) Key Insights:

1. Short Term Prediction is more profitable in Indian stock market as India is a developing country and stock market is greatly influenced by social, political, and company's fundamental factors.
2. Each company under NIFTY-50 Index responded uniquely to various algorithms with distinct technical indicators being important features for different stock companies.
3. It is difficult to postulate a single algorithm that fits well for all the diverse stock portfolios.
4. CNN, which is normally used for Computer Vision, performed well for most of the companies unlike LSTM, MLP and Hybrid Model.
5. Best 5 performing companies in Indian Stock Market were TATA Steel, NTPC, PowerGrid, ITC, and ONGC.

6 C) Study Limitations:

1. Feature Selection: This study did not use dimensionality reduction techniques like Principal Component Analysis for selecting important features.
2. Hyper-parameter Tuning and Model optimization using Gradient descent was not performed for Deep learning models due to large datasets and computational difficulties.
3. This study was subjective to only Indian stock market and was not tested on real-live time trading platform.

6 D) Future Scope:

1. Back-testing strategies for finding profitable gains on short-term predictions can be studied in the future.
2. Algorithmic trading: finding golden standard frameworks and guidelines which are valid and reliable. (CS5704 Week 3 Teaching Materials, 2022) This will help in better understanding of the profitable market trends based on historical stock data which includes factors like price, volume, company's annual reports, technical indicators, suggestions from experts, market capitalization etc.

References:

1. Naik, N. & Mohan, B.R. 2021, "Novel Stock Crisis Prediction Technique-A Study on Indian Stock Market", IEEE access, vol. 9, pp. 1-1.

2. Saji, T.G. 2018, "Financial Distress and Stock Market Failures: Lessons from Indian Realty Sector", Vision (New Delhi, India), vol. 22, no. 1, pp. 50-60.

3. Pattewar, T., Jain, D. & Kiranmayee, B.V. 2023, "A Buffalo-based bi-directional recurrent paradigm for Indian stock market prediction", Concurrency and computation, vol. 35, no. 9.

4. Gadgil, A.S., Fakirmohan Desity, A., Asole, P.H., Shailesh Dandge, H. & Shinde, S. 2021, "Stock Market Prediction through Artificial Intelligence, Machine Learning and Neural Networks", IEEE, , pp. 1.

5. Qiu, M., Song, Y. & Akagi, F. 2016, "Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market", Chaos, solitons and fractals, vol. 85, pp. 1-7.

6. Hiransha, M. et al. (2018b) “NSE Stock market Prediction using Deep-Learning Models,” Procedia Computer Science, 132, pp. 1351–1362. Available at: https://doi.org/10.1016/j.procs.2018.05.050.

7. Weng, B., Ahmed, M.A. & Megahed, F.M. 2017, "Stock market one-day ahead movement prediction using disparate data sources", Expert systems with applications, vol. 79, pp. 153-163.

8. Masoud, N.M.H. 2013, "The impact of stock market performance upon economic growth", International journal of economics and financial issues, vol. 3, no. 4, pp. 788-798.

9. Girish, G.P. (2019). Impact of Implementation of Goods and Services Tax on Nifty 50 Index of National Stock Exchange of India. Theoretical Economics Letters, 09(01), pp.172–179. doi:https://doi.org/10.4236/tel.2019.91014.

10. Aghera, A., Emery, M., Bounds, R., Bush, C., Stansfield, B., Gillett, B. and Santen, S. (2018). A Randomized Trial of SMART Goal Enhanced Debriefing after Simulation to Promote Educational Actions. Western Journal of Emergency Medicine, [online] 19(1), pp.112–120. doi:https://doi.org/10.5811/westjem.2017.11.36524.

11. Shah, D., Isah, H. and Zulkernine, F. (2019) “Stock Market Analysis: A Review and Taxonomy of Prediction Techniques,” International Journal of Financial Studies, 7(2), p. 26. Available at: https://doi.org/10.3390/ijfs7020026.

12. Patel, J. et al. (2015) “Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques,” Expert Systems With Applications, 42(1), pp. 259–268. Available at: https://doi.org/10.1016/j.eswa.2014.07.040.

13. Wang, X. et al. (2022) “Learning nonstationary Time-Series with dynamic pattern extractions,” IEEE Transactions on Artificial Intelligence, 3(5), pp. 778–787. Available at: https://doi.org/10.1109/tai.2021.3130529.

14. Hu, Z., Zhao, Y. & Khushi, M. 2021, "A survey of forex and stock price prediction using deep learning", Applied system innovation, vol. 4, no. 1, pp. 1-30.

15. Grindsted, T.S. (2021) “Algorithmic Finance: Algorithmic Trading across Speculative Time-Spaces,” Annals of the American Association of Geographers, pp. 1–13. Available at: https://doi.org/10.1080/24694452.2021.1963658.

16. Alam, T.M. et al. (2020) “Corporate Bankruptcy Prediction: An approach towards Better Corporate world,” The Computer Journal, 64(11), pp. 1731–1746. Available at: https://doi.org/10.1093/comjnl/bxaa056.

17. Singh, J. and Khushi, M. (2021b) “Feature learning for stock price prediction shows a significant role of analyst rating,” Applied System Innovation, 4(1), p. 17. Available at: https://doi.org/10.3390/asi4010017.

18. Lo, A.W. (2007) Efficient Markets Hypothesis. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=991509.

19. Parmar, I., Agarwal, N., Saxena, S., Arora, R., Gupta, S., Dhiman, H. and Chouhan, L. (2018). Stock Market Prediction Using Machine Learning. 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). doi:https://doi.org/10.1109/icsccc.2018.8703332.

20. Nelson, D.M.Q., Pereira, A.C.M. and de Oliveira, R.A. (2017). Stock market’s price movement prediction with LSTM neural networks. 2017 International Joint Conference on Neural Networks (IJCNN). doi:https://doi.org/10.1109/ijcnn.2017.7966019.

21. Qi, L., Khushi, M. and Poon, J. (2020) “Event-Driven LSTM For Forex Price Prediction,” 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) [Preprint]. Available at: https://doi.org/10.1109/csde50874.2020.9411540.

22. Vijh, M., Chandola, D., Tikkiwal, V.A. and Kumar, A. (2020). Stock Closing Price Prediction using Machine Learning Techniques. Procedia Computer Science, 167(167), pp.599–606. doi:https://doi.org/10.1016/j.procs.2020.03.326.

23. Li, Y., Ni, P. & Chang, V. 2020, "Application of deep reinforcement learning in stock trading strategies and stock forecasting", Computing, vol. 102, no. 6, pp. 1305-1322.

24. Meng, T.L. and Khushi, M. (2019). Reinforcement Learning in Financial Markets. Data, 4(3), p.110. doi:https://doi.org/10.3390/data4030110.

25. Jaggi, M., Mandal, P., Narang, S., Naseem, U. and Khushi, M. (2021). Text Mining of Stocktwits Data for Predicting Stock Prices. Applied System Innovation, [online] 4(1), p.13. doi:https://doi.org/10.3390/asi401001

26. Nti, I.K., Adekoya, A.F. and Weyori, B.A. (2020). Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana. Applied Computer Systems, 25(1), pp.33–42. doi:https://doi.org/10.2478/acss-2020-0004.

27. Gao, T. & Chai, Y. 2018, "Improving Stock Closing Price Prediction Using Recurrent Neural Network and Technical Indicators", Neural computation, vol. 30, no. 10, pp. 2833-2854.

28. Tanaka-Yamawaki, M. & Tokuoka, S. 2007, "Adaptive use of technical indicators for the prediction of intra-day stock prices", Physica A, vol. 383, no. 1, pp. 125-133.

29. The Economic Times. (n.d.). NSE Nifty | Live NSE Nifty Index, NSE India - S&P CNX Nifty. [online] Available at: https://economictimes.indiatimes.com/indices/nifty_50_companies.

30. Zhai, Y., Hsu, A. and Halgamuge, S.K. (2007). Combining News and Technical Indicators in Daily Stock Price Trends Prediction. Advances in Neural Networks – ISNN 2007, pp.1087–1096. doi:https://doi.org/10.1007/978-3-540-72395-0_132.

31. Oncharoen, P. & Vateekul, P. 2018, "Deep Learning for Stock Market Prediction Using Event Embedding and Technical Indicators", IEEE, , pp. 19.

32. Neely, C.J., Rapach, D., Tu, J. and Zhou, G. (2014). Forecasting the Equity Risk Premium: The Role of Technical Indicators. SSRN Electronic Journal. doi:https://doi.org/10.2139/ssrn.1787554.

33. Ma, C. & Yan, S. 2022, "Deep learning in the Chinese stock market: The role of technical indicators", Finance research letters, vol. 49, pp. 103025.

34. Yung-Keun Kwon and Byung-Ro Moon (2007). A Hybrid Neurogenetic Approach for Stock Forecasting. IEEE Transactions on Neural Networks, 18(3), pp.851–864. doi:https://doi.org/10.1109/tnn.2007.891629.

35. Huang, Y., Capretz, L.F. & Ho, D. 2021, "Machine Learning for Stock Prediction Based on Fundamental Analysis", IEEE, Ithaca, pp. 01.

36. Boonpeng, S. & Jeatrakul, P. 2014, "Enhance the performance of neural networks for stock market prediction: An analytical study", IEEE, pp. 1.

37. ZHAO, X. & ZHAO, Q. 2021, "Stock Prediction Using Optimized LightGBM Based on Cost Awareness", IEEE,pp. 107.

38. Gómez-Escalonilla, V., Martínez-Santos, P. & Martín-Loeches, M. 2022, "Preprocessing approaches in machine-learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali", Hydrology and earth system sciences, vol. 26, no. 2, pp. 221-243.

39. GRIGORYAN, H. 2016, "A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)", Database systems journal, vol. VII, no. 1, pp. 12-21.

40. Shynkevich, Y., McGinnity, T.M., Coleman, S.A., Belatreche, A. and Li, Y. (2017). Forecasting price movements using technical indicators: Investigating the impact of varying input window length. Neurocomputing, 264, pp.71–88. doi:https://doi.org/10.1016/j.neucom.2016.11.095.

41. Jain, S.L., Gupta, R. and Moghe, A.A. (2018). Stock Price Prediction on Daily Stock Data using Deep Neural Networks. 2018 International Conference on Advanced Computation and Telecommunication (ICACAT). doi:https://doi.org/10.1109/icacat.2018.8933791.

42. Althelaya, K.A., El-Alfy, E.-S.M. and Mohammed, S. (2018). Evaluation of bidirectional LSTM for short- and long-term stock market prediction. 2018 9th International Conference on Information and Communication Systems (ICICS). doi:https://doi.org/10.1109/iacs.2018.8355458.

43. Narayana Darapaneni, Anwesh Reddy Paduri, Sharma, H., Manjrekar, M., Hindlekar, N., Bhagat, P., Aiyer, U. and Agarwal, Y.K. (2022). Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2204.05783.

44. Rajab, S. and Sharma, V. (2015). Performance evaluation of ANN and Neuro-fuzzy system in business forecasting. [online] IEEE Xplore. Available at: https://ieeexplore.ieee.org/abstract/document/7100349.

45. Hoque, K.E. and Aljamaan, H. (2021). Impact of Hyperparameter Tuning on Machine Learning Models in Stock Price Forecasting. IEEE Access, pp.1–1. doi:https://doi.org/10.1109/access.2021.3134138.

46. Fathali, Z., Kodia, Z. and Ben Said, L. (2022). Stock Market Prediction of NIFTY 50 Index Applying Machine Learning Techniques. Applied Artificial Intelligence, 36(1). doi:https://doi.org/10.1080/08839514.2022.2111134.

47. Sisodia, P.S., Gupta, A., Kumar, Y. and Ameta, G.K. (2022). Stock Market Analysis and Prediction for Nifty50 using LSTM Deep Learning Approach. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICIPTM54933.2022.9754148.

48. Vikalp Ravi Jain, Gupta, M. and Raj Mohan Singh (2018). Analysis and Prediction of Individual Stock Prices of Financial Sector Companies in NIFTY50. International Journal of Information Engineering and Electronic Business, 10(2), pp.33–41. doi:https://doi.org/10.5815/ijieeb.2018.02.05

49. Selvamuthu, D., Kumar, V. & Mishra, A. 2019, "Indian stock market prediction using artificial neural networks on tick data", Financial innovation (Heidelberg), vol. 5, no. 1, pp. 1-12.

50. Raviraj, S., M M, M.P. & Pai, K.M. 2021, "Share price prediction of Indian Stock Markets using timeseries data - A Deep Learning Approach", IEEE, , pp. 744.

51. Banerjee, S. & Mukherjee, D. 2022, "Short Term Stock Price Prediction in Indian Market: A Neural Network Perspective", Studies in microeconomics, vol. 10, no. 1, pp. 23-49.

52. Han, Y., Wu, J., Zhai, B., Pan, Y., Huang, G., Wu, L. and Zeng, W. (2019). Coupling a Bat Algorithm with XGBoost to Estimate Reference Evapotranspiration in the Arid and Semiarid Regions of China. Advances in Meteorology, 2019, pp.1–16. doi:https://doi.org/10.1155/2019/9575782.

53. Nabipour, M. et al. (2020) “Deep learning for stock market prediction,” Entropy, 22(8), p. 840. Available at: https://doi.org/10.3390/e22080840

54. Sharma, N., Soni, M., Kumar, S., Kumar, R., Deb, N. &amp; Shrivastava, A. 2023, &quot; Supervised Machine Learning Method for Ontology-based Financial Decisions in the Stock Market&quot;, ACM transactions on Asian and low-resource language information processing, vol. 22, no. 5, pp. 1-24.

55. Malladi, R.K. (2022) “Application of supervised machine learning techniques to forecast the COVID-19 U.S. recession and stock market crash,” Computational Economics [Preprint]. Available at: https://doi.org/10.1007/s10614-022-10333-8.

56. Yelne and D. Theng, "Stock Prediction and analysis Using Supervised Machine Learning Algorithms," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), Nagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCICA52458.2021.9697162.

57. Subasi, A. et al. (2021) “Stock market prediction using machine learning,” Procedia Computer Science, 194, pp. 173–179. Available at: https://doi.org/10.1016/j.procs.2021.10.071.

58. Kumar, D., Sarangi, P.K. and Verma, R. (2022) “A systematic review of stock market prediction using machine learning and statistical techniques,” Materials Today: Proceedings, 49, pp. 3187–3191. Available at: https://doi.org/10.1016/j.matpr.2020.11.399.

59. Iyyappan, M. et al. (2022) “A novel AI-Based stock market prediction using machine learning algorithm,” Scientific Programming, 2022, pp. 1–11. Available at: https://doi.org/10.1155/2022/480808

60. CS5704 Week 1 Teaching Materials (2022). Lecture 1: Dissertation Planning. Available at https://brightspace.brunel.ac.uk, accessed 11 September 2023

61. CS5704 Week 3 Teaching Materials (2022). Lecture 3: Research Methodology. Available at https://brightspace.brunel.ac.uk, accessed 11 September 2023

62. CS5704 Week 6 Teaching Materials (2022). Lecture 6: Project Planning an Estimation. Available at https://brightspace.brunel.ac.uk, accessed 11 September 2023

Appendix:

Ethical Approval Letter:

College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

21 July 2023

LETTER OF CONFIRMATION

Applicant: Ms Drishti Doshi

Project Title: Predicting the Indian Stock Market using NIFTY-50 Index

Reference: 43842-NER-Jun/2023- 45826-1

Dear Ms Drishti Doshi,

The Research Ethics Committee has considered the above application recently submitted by you.

This letter is to confirm that, according to the information provided in your BREO application, your project does not require full ethical review. You may proceed with your research as set out in your submitted BREO application, using secondary data sources only. You may not use any data sources for which you have not sought approval.

Please note that:

- **You are not permitted to conduct research involving human participants, their tissue and/or their data. If you wish to conduct such research (including surveys, questionnaires, interviews etc.), you must contact the Research Ethics Committee to seek approval prior to engaging with any participants or working with data for which you do not have approval.**
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must submit a new BREO application and await approval before proceeding. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Good luck with your research!

Kind regards,

Simon Taylor

Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee Brunel

University London