

Stock Trend prediction

Submitted By

Deep Metaliya

22MCED10



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INSTITUTE OF TECHNOLOGY
NIRMA UNIVERSITY
AHMEDABAD-382481**

May 2024

Stock Trend prediction

Major Project - II

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering (Data Science)

Submitted By

Deep Metaliya

(22MCED10)

Guided By

Dr. Ankit Thakkar



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

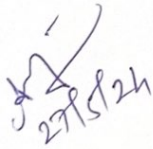
NIRMA UNIVERSITY

AHMEDABAD-382481

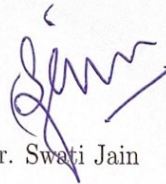
May 2024

Certificate

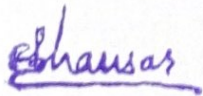
This is to certify that the major project entitled “**Stock Trend prediction**” submitted by **Deep Metaliya (Roll No : 22MCED10)** towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering (Data Science) of Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project part-II, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.



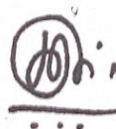
Dr. Ankit Thakkar
Guide & Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.



Dr. Swati Jain
Associate Professor,
Coordinator M.Tech - DS
Institute of Technology,
Nirma University, Ahmedabad



Dr. Madhuri Bhavsar
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.



Dr. Himanshu Soni
Director,
School of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, Deep Metaliya, Roll. No. 22MCED10, give undertaking that the Major Project entitled "Stock Trend prediction" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science and Engineering (Data Science)** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.



Signature of Student

Date: 27/5/24

Place: Ahmedabad



Endorsed by

Dr. Ankit Thakkar

(Signature of Guide)

Acknowledgements

I take great pleasure in expressing my heartfelt thanks and profound gratitude to **Dr. Ankit Thakkar**, Professor in the Computer Engineering Department at the Institute of Technology, Nirma University, Ahmedabad. His invaluable guidance and unwavering encouragement have been instrumental in the success of this endeavor. The continuous support and motivation he provided have propelled me towards achieving higher goals, and I am truly grateful for the intellectual maturity nurtured under his mentorship, which will undoubtedly benefit me in the long run.

I also extend my sincere appreciation to **Dr. Madhuri Bhavsar**, Head of the Computer Science and Engineering Department at the Institute of Technology, Nirma University, Ahmedabad. I am thankful for her kind support and for fostering a conducive research environment by providing essential infrastructure.

A special note of gratitude goes to **Dr. Himanshu Soni**, the Hon'ble Director of the Institute of Technology, Nirma University, Ahmedabad. His unparalleled motivation has been a driving force throughout the course of this work, and I am truly appreciative of his unwavering support.

I would like to express my thanks to the entire institution and all the faculty members of the Computer Engineering Department at Nirma University, Ahmedabad, for their special attention and valuable suggestions that contributed to the success of the project.

Deep Metaliya

22MCED10

Abstract

The rapid evolution of financial markets demands advanced tools and methodologies for effective decision-making. This project explores the area of stock trend prediction by leveraging the power of deep learning techniques. It is aimed at developing a reliable and nearby forecast model that will help Bulk Investors, Business people, Retail investors and Economic analysts to make some decisions.

Our approach involves the utilization of deep learning algorithms, specifically neural networks, to analyze historical stock market data and extract meaningful patterns. To improve the accuracy and make effective model we also add some calculated features to the model. The project employs a comprehensive dataset encompassing diverse financial indicators and historical stock prices to train the neural network model.

Key components of the project include data preprocessing, feature selection, model architecture design and training/validation strategies. Evaluation will be done by testing on seen data.

Abbreviations

Stock Market	Financial market for trading securities.
SMT	Stock Market Trend.
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
K-means	K-means algorithm
MR	Median Range
FS	Feature Selection
GA	Genetic Algorithm
ICA	Independent Components Analysis

—

Contents

Certificate	iii
Statement of Originality	iv
Acknowledgements	v
Abstract	vi
Abbreviations	vii
List of Tables	x
List of Figures	xi
1 Introduction	1
1.1 Basic Introduction	1
1.2 Objective	2
1.3 Problem Statement	2
1.4 Scope of Investigation	2
2 Literature Survey	3
3 Methodology	8
3.1 Data collection	8
3.2 Cleaning	9
3.3 Additional features	9
3.4 Normalization	17
3.5 Coefficient of Variation	17
3.6 Feature selection Methods for calculated CV	18
3.6.1 K-means algorithm	18
3.6.2 Median Range	18
3.6.3 Top-M	18
3.7 Feature Selection method via ICA and Genetic Algorithm	19
3.7.1 Independent Components Analysis	19
3.7.2 Genetic Algorithm	19
3.8 Convolutional Neural Network (CNN) Architecture	20
3.8.1 Training Configuration	20
3.8.2 Dataset Splitting	20
3.9 Parameter for GA-LSTM model	21
3.9.1 Genetic Algorithm	21

3.9.2	LSTM	21
3.10	Parameter of LSTM model for feature selection using ICA	22
3.10.1	LSTM	22
4	Performance evaluation	25
4.1	R2 Score (Coefficient of Determination)	25
4.2	Mean Squared Error (MSE)	26
4.3	Mean Absolute Error(MAE)	27
4.4	Top-M	28
4.5	Cluster	29
4.6	Median Range	30
4.7	Independent Components Analysis(ICA)	31
4.8	GA-LSTM	32
5	Conclusion remarks and future directives	33
5.1	Summary	33
5.2	Conclusion	34
5.3	Future Directive	34
	Bibliography	35

List of Tables

2.1 Literature survey table	6
---------------------------------------	---

List of Figures

3.1	Flow of the Existing approach for feature selection[1]	23
4.1	R2 score for M Cluster	28
4.2	Execution Time For Top M Cluster	28
4.3	R2 score for Top M Feature lies between Median Range	29
4.4	Execution Time For Top M feature Lies Between Median Range	29
4.5	R2 score for Top M Feature	30
4.6	Execution Time For Top M feature	30
4.7	accuracy comparison for 48 Features using ICA (LSTM)	31
4.8	accuracy comparison for 48 Features using ICA (GRU)	31
4.9	MSE comparison for 40 Features using GA-LSTM(CCB)	32
4.10	MSE comparison for 40 Features using GA-LSTM(CSI 300)	32

Chapter 1

Introduction

1.1 Basic Introduction

Stock market trends are difficult to predict in the fast-paced world of finance. Investors and experts are always trying to figure them out. Being able to predict how the market will move can make a big difference in making smart decisions and getting the best results. Traditional methods have their limits, which is why people are always looking for better ones. In this report, we look how we can take advantage of deep learning models, especially CNN, LSTM and GRU can help us figure out how to predict stock market trends.

To be able to see the subtleties of market operations, you need both traditional models and modern methodologies for picking out features that are useful. This study looks at a number of methods, such as the Coefficient of Variation (CV) to find relevant features, the K-means algorithm to group features that are similar, and the Median Range (MR) to deal with outliers and we also used ICA and GA for Features selection. We want to improve the accuracy and dependability of our predictions by using these methods.

We will talk a lot about feature selection as we talk about the importance of CV value, the grouping power of K-means, and the Median Range's ability to find outliers. Together, these methods make the model more stable and well-tuned.

Next, we'll talk about the structure of CNN model which used to predict stock market trends and the steps used to train it and we also talk about LSTM and GRU for prediction. We set evaluation metrics that check the model's performance in terms of accuracy MSE, MAE and R2 Score to see how well our method works.

1.2 Objective

The main Focus of this study is to use advanced deep learning methods, especially CNN, LSTM and GRU to make predicting stock market trends more accurate and useful. The study aims to fix the problems with current methods by using new ways to choose features that are based on the Coefficient of Variation value, the K-means algorithm, the Median Range, ICA and Genetic Algorithm.

1.3 Problem Statement

The most important problem at hand is making accurate predictions about how the stock market will move in the coming months or days. Existing models and methods have a hard time keeping up with how the market is changing, and they might not be making the best use of the amount of information that is stored in financial data. The problem becomes even more acute because of the noise and randomities that are inherent in market data. Therefore, in order to enhance models' reliability it is important that the methods of selecting features are improved.

1.4 Scope of Investigation

This study use advanced deep learning methods, Mostly CNN, LSTM, GRU to predict stock market trends. The study's goal is to find out how well models can find complicated patterns and connections in financial data so that trend estimates are more accurate.

In addition, the study looks into different ways to choose features, such as using the Coefficient of Variation (CV) value, the K-means algorithm for grouping features, and the Median Range (MR) for finding and dealing with outliers. The model is better able to find important features and is more stable when it comes to dealing with noisy and irregular data because of these techniques.

The scope also provides an opportunity to provide insight and methodologies which can improve the accuracy and reliability of stock market trend forecasting through potential contributions from this study in the field of finance analysis.

Chapter 2

Literature Survey

It suggests the MS-SSA-LSTM model for predicting stock prices. This model combines deep learning, the Sparrow Search Algorithm (SSA) for optimising LSTM hyperparameters, and mood analysis. It works better than other models; compared to normal LSTM, it has an R-squared improvement of 10.74%[\[2\]](#). Adding sentiment analysis improves the model's ability to make predictions, SSA makes the LSTM hyperparameters work better, and the model works especially well for making short-term predictions in China's volatile financial market.

A new bio-inspired method called Artificial Rabbits Optimisation (ARO), which is based on how rabbits survive in the wild. The algorithm uses random hiding to stay out of the way of predators and reroute foraging to protect nests. By modelling these methods mathematically, we can make an optimizer that works well. It is clear that ARO works because it beats other optimizers at solving test functions and engineering problems[\[3\]](#). In addition, ARO has proven to be very effective in enhancing the backpropagation networks that are able to detect faults on rolling bearings and showing how helpful it is for a realistic environment.

The SA-DLSTM hybrid model, which is a mix of an Emotion enhanced CNN (ECNN), denoising autoencoder (DAE) and a Long Short-Term Memory (LSTM) model. It is meant to make stock market predictions easier. ECNN gets a sense of how people feel by using notes left by internet users as extra data, and DAE improves important parts of stock market data to make predictions more accurate[\[4\]](#). When making realistic mood indexes, the speed with which emotions change on the stock market is taken into account.

When you combine key features and mood indexes into LSTM, you get more accurate predictions than with other models. This shows that it works well for both return and risk, which helps investors make smart decisions.

Machine learning-based modelling methods to help with the difficulties of guessing stock prices and finds signals in candlestick pattern in charts. For correct stock price predictions, a Vector Auto regression based rolling prediction model is put forward. A Gaussian Feed-Forward NN method is also shown for graphic signal identification, which makes it easier to spot different stock price signs. The results of the experiments show that these methods work better than current ones[5]. This proves that they can be used in real-life stock exchange strategies to help people make smart investment decisions.

In another research it focuses on well-known off-the-shelf models such as ANN, SVM, Random Forest (RF) and Naive-Bayes (NB). The study shows that these models are not good at predicting the direction of next day's ending prices, by comparing the results to predictions for the current day. The study questions whether these models can be used in the real world by showing that, even though they are good at predicting what will happen today, they are about as good as guessing when it comes to predicting short-term market trends for the future[6]. The results add more support to the Efficient Market Hypothesis and add to the conversation about algorithmic trade.

Machine learning models have been used a lot in quantitative stock trading systems that are driven by data and technology to predict how stocks will react the next day. This study looks at the problems that come up when you try to use information from Japanese candlestick charts in trade systems that are based on machine learning[7]. Here research suggests that machine learning and pattern recognition steps to help reduce the problem of too many false signals. It says to pick and choose which trading suggestions might not be reliable based on known graphic trends. Different mixes of pattern recognition methods are looked into, including shallow and deep supervised models, as well as autoregressive methods. The method works well in terms of return on investment, as shown by results from experiments done in various market exchanges and situations.

The stock market is a very important sign of a country's economic health because it shows how well businesses are doing generally by buying and selling shares of publicly

traded companies. Investing has risks, but it could pay off in the long run. It has become more popular to use artificial intelligence (AI) in the financial field, specially LSTM for time series analysis. A research looks at how the Artificial Rabbits Optimisation (ARO) method can be used to improve the accuracy of stock market predictions by making the LSTM hyper-parameters better. The study shows that an improved deep LSTM network with ARO (LSTM-ARO) is better at predicting stock prices than other models[8], such as ANN, different LSTM setups and LSTM optimised by the Genetic Algorithm. MSE, MAE, Mean Absolute Percentage Error (MAPE), and R-squared (R2) are some of the tests that show that LSTM-ARO is better than its competitors.

New Deep Learning techniques are emerging as technology continues to change rapidly. These techniques have a number of uses in finance, such as forecasting the equity markets, optimising portfolios, managing risks and developing trading strategies. The main focus of this study is on predicting the DAX, DOW, and S&P500 stock markets while taking into account the complexity of noisy data. The use of recurrent neural network based models, such as CNN-LSTM, GRUWNTANT and ensembles is suggested for new mixed models. A unique feature is taking the average of the stock market measures' high and low prices. The results of the experiments show that that models work better than traditional machine-learning models in 81.5% of cases for more time frame forecasting, 48.1% of cases for one time frame forecasting, and 40.7%[9] of cases for traditional machine-learning models.

Financial researchers have long focused on predicting stock trends. Despite the large quantity of information accessible now, past attempts to estimate trends using textual data were limited due to fixed word embeddings and dependence on general market sentiment. This work presents a deep learning model for predicting trends in the Thailand Futures Exchange (TFEX) by analysing numerical and textual data. Thai economic news stories are categorised into industry-specific indices to better reflect sector-specific fluctuations. The suggested technique predicts daily stock market activity using Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) architectures.[1] The experimental findings suggest that combining both numerical and textual information at the sector level improves prediction accuracy and exceeds current baselines.

Table 2.1: Literature survey table

Paper title	Published year	Dataset used	Algorithms / Methods used	Pros of paper/ novel work	Cons of paper / Limitations
Shortlisting machine learning-based stock trading recommendations using candlestick pattern recognition[7]	2023	S&P 500, FTSE MIB40	LSTM, ARIMA, ExpSmooth, Candlestick pattern recognition	Filters ML based trading recommendations that are unreliable according to recognized graphical patterns	Classifier training and computation time in intensive phase is high.
Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm[8]	2023	DJIA stock dataset	Optimized deep LSTM, ANN, GA, ARO	Novel LSTM network is created with parameters	Static trading system, lack of real-time analysis.
Deep Learning-based Integrated Framework for stock price movement prediction[4]	2023	Daily trading data of HSI	LSTM, SA-DLSTM, Sentiment analysis	Novel hybrid model SA-DLSTM	Less efficient to process voluminous data in real-time.
Stock Price Prediction Model Based on Investor Sentiment and Optimized Deep Learning[2]	2023	China's A-share market	LSTM, SPARROW SEARCH ALGORITHM	Combination of Sentiments and Technical	News data should be over Change with time
Machine learning techniques for stock price prediction and graphic signal recognition[5]	2023	randomly chose several stocks	VAR, GFNN	Combination of Graphical pattern and Machine learning	Pattern will not be the exact same every time

Forecasting Stock Market Indices Using the Recurrent Neural Network Based Hybrid Models: CNN-LSTM, GRU-CNN, and Ensemble Models[9]	2023	DAX,DOW,S&P500	CNN, LSTM, RNN, Ensemble	Easy to understand, no need of extra data	Results might be change with different parameters
Stock Trend Prediction Using Deep Learning Approach on Technical Indicator and Industrial Specific Information[10]	2021	SET50	LSTM, Bidirectional Encoder Representations from Transformers (BERT)	Improved performance, achieved highest annualized return,	highly dependent on the quality and relevance of the industry-specific information

Chapter 3

Methodology

In this implementation we used some advance deep learning technique like CNN, LSTM and GRU here, to figure out how to predict stock market trends in a very complicated world. Our method includes advanced feature selection based on Coefficient of variation (CV). We used K-means algorithm for grouping similar features and the Median Range (MR) for finding outliers. The goal is to improve the accuracy of our predictions. And we also use ICA and GA to select useful features. In this part we describes the steps that were taken to collect data in a structured way, normalized it, as well as the layout of the CNN, LSTM and GRU model that was made to predict stock market trends. We provide a flow for the use of new techniques that aim to simplify the complicated world of financial markets by explaining the approach.

3.1 Data collection

- Utilized the yfinance python library to access historical stock market data.
- For time frame from March 1, 2000 to October 30, 2017 data set taken for S&P 500 index to analysis[1][11].
- And we data data for China construction bank, CSI 300 stock for time frame of January 1, 2010 to April 1, 2020[12]

3.2 Cleaning

- To improve the integrity of the data we removed null values and blank values from the time series data.
- And for calculated features we put mean values where null values are there into the dataframe.

3.3 Additional features

Along with the data we take from yfinance we add some following features that were designed and added to the data frame to improve the time series data set's analysis capabilities.

1. MACD (Moving Average Convergence Divergence)

- The moving average convergence/divergence (MACD) or MAC-D indicator shows how two EMA of a Stock price are related to each other. It is a trend-following momentum indicator[13].
- Here we are calculating 14 and 21 days MACD to further use in future.

2. Return

- Return is a percentage change in the closing price from previous day[13].
- We are also calculating Past 8 Weekly & Past 2 Monthly Returns

3. Pivot Point

- A pivot point is a technical analysis indication or calculation that is used to assess the market's overall trend throughout various time periods. The pivot point is just the average of the intraday high and low, as well as the preceding trading day's closing price[13].

$$P = \frac{H + L + C}{3}$$

where:

H = High price of the previous day

L = Low price of the previous day

C = Closing price of the previous day

4. Momentum

- Momentum is the rate at which the price of a stock, asset, or other tradable object changes. Momentum is a way for buyers to figure out how strong a trend is by showing them how fast prices change over time[13].

The momentum of a stock is calculated as follows:

$$\text{Momentum} = \text{Price}_{\text{today}} - \text{Price}_{n \text{ days ago}}$$

where:

$\text{Price}_{\text{today}}$ = Closing price of today

$\text{Price}_{n \text{ days ago}}$ = Closing price n days ago

5. Average True Range (ATR))

- A true range is calculated by the difference between the current high, current low and close from the previous bar. This average of a true range over the specified period is known as Average True Range (ATR). ATR measures volatility, extending the price movement beyond any gaps or skips. Normally, the stipulations of the equation that determine the ATR are done with 14 periods in consideration that can be end-of-period intraday, for example, or daily, weekly, or monthly[13].

$$TR = \max(\text{High} - \text{Low}, |\text{High} - \text{Previous Close}|, |\text{Low} - \text{Previous Close}|)$$

$$ATR = \frac{1}{n} \sum_{i=1}^n TR_i$$

where:

High = Highest price of current period

Low = Lowest price of current period

Previous Close = Closing price of previous period

n = Number of periods

6. SMA (Simple Moving Average)

- Computed Simple Moving Average to smooth out price fluctuations and identify trends over a specified period. It is a average value of previous specific period value[13].

7. Exponential MA

- An exponential moving average(EMA) gives the newest data points more weight and importance than older ones[13].

8. RSI (Relative Strength Index)

- Technical analysis uses the relative strength index (RSI) to show how fast prices are moving. RSI checks to see if the price of an investment is overvalued or cheap by looking at how quickly and how much it has changed recently[13].
- Relative Strength Index (RSI) Calculation

Step 1: Calculate Average Gain and Average Loss

$$\text{AvgGain} = \frac{\text{Sum of gains over the last 14 days}}{14}$$

$$\text{AvgLoss} = \frac{\text{Sum of losses over the last 14 days}}{14}$$

Step 2: Calculate Relative Strength (RS)

$$\text{RS} = \frac{\text{AvgGain}}{\text{AvgLoss}}$$

Step 3: Calculate RSI

$$\text{RSI} = 100 - \frac{100}{1 + \text{RS}}$$

where:

AvgGain = Average gain over the specified period (typically 14 days)

AvgLoss = Average loss over the specified period (typically 14 days)

RS = Relative Strength

RSI = Relative Strength Index

9. Bollinger Bands

- Bollinger Band widths are computed as two standard deviations from the price chart's Simple Moving Average, which represents an uptrend or downtrend. As a result, the width of the bands is calculated as a standard deviation. The spreads automatically adjust to movements in the underlying price. The Bollinger Bands are based on two parameters: time and standard deviations, the latter of which is known as StdDev[13].

- Middle Band (MB)

$$MB = SMA_n$$

- Upper Band (UB)

$$UB = MB + (k \times \sigma_n)$$

- Lower Band (LB)

$$LB = MB - (k \times \sigma_n)$$

where:

SMA_n = n-period simple moving average

σ_n = n-period standard deviation

k = number of standard deviations

10. OBV (On-Balance Volume)

- On-balance volume shows whether trading volume is going into or out of a certain investment or currency pair and keeps track of the total trading volume for that asset. The OBV is the sum of all the volumes, both positive and negative[13].

11. Stochastic Oscillator

- A stochastic oscillator is a moving average or momentum indicator which compares a specific closing price of a security to a band the set prices move within over a certain time period. Such oscillator is sensitive to market forces, but this sensitivity can be reduced by changing the period of time or by averaging the moving result. It is meant to create buy and sell signals that are based on the trading momentum by using range of values which is bounded with 0-100[13].

The Stochastic Oscillator is calculated as follows:

- %K Line

$$\%K = \frac{\text{Current Close} - \text{Lowest Low}}{\text{Highest High} - \text{Lowest Low}} \times 100$$

- %D Line

$$\%D = \text{SMA}_3(\%K)$$

where:

Current Close = Most recent closing price

Lowest Low = Lowest low over the look-back period

Highest High = Highest high over the look-back period

$\text{SMA}_3(\%K)$ = 3-period simple moving average of the %K Line

12. Trending Fibonacci Retracement

- Technical analysis assumes that Fibonacci retracement levels are the places in which a stock can pause or resume at the moment the price has been reversed. Common ratios include 23.6%, 38.2%, and 50% percent are numbered among these. This scenario of these signals would mostly happen when sequences of simultaneous high and low of the security price are used as determinants of their future[13].

Fibonacci retracement levels are calculated as follows :

- Uptrend Fibonacci Retracement Levels

In an uptrend, the retracement levels are:

38.2% Retracement Level

$$\text{Retracement}_{38.2} = H - 0.382 \times (H - L)$$

50% Retracement Level

$$\text{Retracement}_{50} = H - 0.5 \times (H - L)$$

61.8% Retracement Level

$$\text{Retracement}_{61.8} = H - 0.618 \times (H - L)$$

- Downtrend Fibonacci Retracement Levels

In a downtrend, the retracement levels are:

38.2% Retracement Level

$$\text{Retracement}_{38.2} = L + 0.382 \times (H - L)$$

50% Retracement Level

$$\text{Retracement}_{50} = L + 0.5 \times (H - L)$$

61.8% Retracement Level

$$\text{Retracement}_{61.8} = L + 0.618 \times (H - L)$$

where:

H = High point of given period

L = Low point of given period

13. ROC (Rate of Change)

- The Price Rate of Change (ROC) is a momentum-based technical indicator that shows how much the price has changed since it was last seen a certain number of times ago[13].

The Rate of Change (ROC) is calculated as follows:

$$ROC = \frac{\text{Current Price} - \text{Price}_n \text{ periods ago}}{\text{Price}_n \text{ periods ago}} \times 100$$

where:

Current Price = Recent closing price

Price _{n} periods ago = Closing price n periods ago

n = Number of periods

14. Signal Line

- Determined the Signal Line for some period EMA of the MACD to generate buy or sell signals in conjunction with MACD crossovers[13].

These extra features add up to a full set of indicators that make it possible to analyse the S&P 500 time series data in more detail and find different market trends and possible signs.

3.4 Normalization

Normalisation is an important step in getting the information ready for strong analysis. Scaling numerical features within a consistent range gets rid of errors and makes sure that comparisons between variables are fair.

MinMax Scaling - We Used MinMax scaling to turn numbers like closing prices and trading volumes into a standard range. This method keeps the relative relationships between data points and reduce the effect of outlier.

Z-Score Standardisation - A Z-Score, also known as a standard score, is a method for standardising scores on the same scale by dividing a score's variance by the normal deviation in a dataset. The result is a standard score. It calculates the number of standard deviations a particular data point is from the mean. Z-scores can be negative or positive.

3.5 Coefficient of Variation

One way to look at the relative variability of a collection is to use the coefficient of variation. This number tells you how spread out the values are. The coefficient of variation was used to figure out how volatile key financial metrics were in this time series study of the S&P 500 from March 1, 2000, to October 30, 2017.

Calculation Formula - The CV is calculated as the ratio of the standard deviation to the mean[14].

Mathematically,

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}}$$

3.6 Feature selection Methods for calculated CV

The CV values of the derived features are used to choose which features to use. We combine known and new methods to make a set of features from the ones that are already out there. These methods can be used for feature selection based on CV. We want to choose M features so that $1 \leq M \leq N$. Here are some details about how to choose M features.

3.6.1 K-means algorithm

Data point in each cluster is used to the efficiency of the clusters. Find optimal number for cluster using silhouette coefficient [15]. The biggest cluster is picked during feature selection, which is in line with the goal of recording unique patterns and traits in the dataset, which leads to useful cluster assignments and makes it easier to choose features based on the clusters that have been found.

3.6.2 Median Range

To ensure the median value, we arrange the CV data points in order and compute the median, represented as m . Creating a range centered around the median involves subtracting half of the median value and adding it to the median itself, establishing the starting and ending points of the range, respectively. This yields the median range $[m - m/2, m + m/2]$. By examining CV data points within this range, we pinpoint the associated features in the process of feature selection.

3.6.3 Top-M

As we already said, a bigger coefficient of variation (CV) means that features are more likely to be different from one another. Having features that vary a lot can make the learning process of a prediction model a lot better. In this case, a different way to choose features is suggested. It's called "Top-M selection," and it sorts data points in decreasing order based on their CV values. Then, the top features with the highest CV values are picked out to be looked at in more detail.

3.7 Feature Selection method via ICA and Genetic Algorithm

3.7.1 Independent Components Analysis

Independent Component Analysis (ICA) is a statistical and computational approach used in machine learning and deep learning to break down a multivariate signal into distinct non-Gaussian components. The purpose of ICA is to discover a linear transformation for the data that is as close to statistical independence as feasible[11].

The idea of statistical independence is central to the ICA approach. ICA identifies components of mixed signals that are statistically independent of one another.

3.7.2 Genetic Algorithm

Genetic Algorithm (GA) represents an adaptive heuristic search algorithm, which draws its roots and inspiration from the natural selection and genetic evolution theory. It is commonly used to solve optimization problems with large turnout space by approaching the most optimized solution, while it can also be used to select the features in this process. Ruled by a genetic algorithm, the individuals that compose the social structure of the population are represented by a string of chromosomes which the optimization procedure searches for the best solution[12]. Chromosomes, such as consisting of multiple genes, play a self-regulated role by the internal determination of the external presentation of individual traits.

3.8 Convolutional Neural Network (CNN) Architecture

To predict the Adjusted Close Price in the stock market, a CNN model is implemented. The architecture consists of five layers with the following configurations: 64 filters in the first layer, 128 filters in the second layer, 256 filters in the third layer then 512 filters in the fourth layer and a final output layer with 1 neuron for predicting the target Adjusted Close Price[1].

3.8.1 Training Configuration

The bellow given parameter are taken from the paper, ‘Neural network systems with an integrated coefficient of variation-based feature selection for stock price and trend prediction’[1].

- **Epochs:** The model is trained for 125 epochs, ensuring an adequate number of iterations for learning patterns within the data as suggested in paper.
- **Activation Function:** The activation function used throughout the CNN layers is linear, suitable for regression tasks where predicting a continuous value is required.
- **Optimizer:** The Adam optimizer is employed to efficiently adjust the model’s weights during training, optimizing the learning process.
- **Loss Function:** MSE is chosen as the loss function, enabling the model to reduce the difference between predicted and actual Adjusted Close Price values.

3.8.2 Dataset Splitting

The dataset is split into training (3053 records), validation (525 records), and testing (848 records) sets. This ensures a robust evaluation of the model’s performance on seen data.

3.9 Parameter for GA-LSTM model

The bellow given parameter are taken from the paper ‘Stock prediction based on genetic algorithm feature selection and long short-term memory neural network’[12].

3.9.1 Genetic Algorithm

- **Population size:** We are 100 as a initial population.
- **Crossover rate:** Here we are taking crossover rate as 0.8.
- **Mutation rate:** That is the possibility that a single bit would be flipped to the individual’s chromosomes (0.003 here). A mutation implies the creation of the fresh genetic material in the population.
- **Chromosome size:** Chromosome size is a total number of features as we have 40feature in our dataframe.
- **Iterations:** This is the ensuing number of generations that GA will be running (100 generations in the case).

3.9.2 LSTM

- **Epochs:** The model is trained for 100 epochs, ensuring an adequate number of iterations for learning patterns within the data as suggested in paper.
- **Activation Function:** The activation function used throughout the LSTM layers is Elu, suitable for regression tasks where predicting a continuous value is required.
- **Optimizer:** The Adam optimizer is employed to efficiently adjust the model’s weights during training, optimizing the learning process.
- **Loss Function:** MSE is chosen as the loss function, enabling the model to reduce the difference between predicted and actual Adjusted Close Price values.
- **Dataset Splitting:** The dataset is split into training and testing into 8:2 ratio. This ensures a robust evaluation of the model’s performance on seen data.

- **Layers:** The model includes three network layers: input, hidden, and output. The hidden and output layers contain 128 and 1 neurons, respectively.

3.10 Parameter of LSTM model for feature selection using ICA

The bellow given parameter are taken from the paper ‘Application of LSTM, GRU and ICA for stock price prediction’[11].

3.10.1 LSTM

- **Epochs:** The model is trained for 125 epochs, ensuring an adequate number of iterations for learning patterns within the data as suggested in paper.
- **Activation Function:** The activation function used throughout the LSTM layers is linear, suitable for regression tasks where predicting a continuous value is required.
- **Optimizer:** The Adam optimizer is employed to efficiently adjust the model’s weights during training, optimizing the learning process.
- **Loss Function:** Root mean squared error is chosen as the loss function, enabling the model to reduce the difference between predicted and actual Adjusted Close Price values.
- **Dataset Splitting:** The dataset is split into training and testing into 8:2 ratio. This ensures a robust evaluation of the model’s performance on seen data.

This LSTM model consists of five layers where two recurrent layers with 64 and 128 units, two dense levels with 256 and 512 units, and a single output layer used.

As shown in figure 3.1 here are 3 main steps of the model which are followed for CV based method

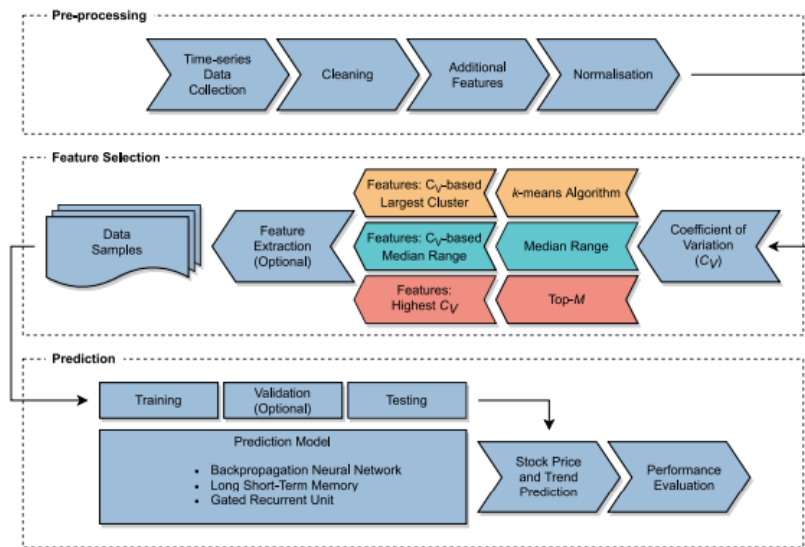


Figure 3.1: Flow of the Existing approach for feature selection[1] .

- **Preprocessing**

- Data Collection
- Cleaning
- Additional Features
- Normalization

- **Feature selection**

- Calculate CV
- Feature Extraction
- Data Sample

- **Prediction**

- Training, Validation, Test
- Evaluation

The conceptual clarity and systematic process of CNN model provided by the flow diagram for feature selection based on calculated CV value, is attributed to the work presented in the research paper, ‘Neural network systems with an integrated coefficient of variation-based feature selection for stock price and trend prediction’[\[1\]](#).

Chapter 4

Performance evaluation

4.1 R2 Score (Coefficient of Determination)

In a regression model, R-squared (R2) is a statistical measure that shows how much of the variation in a dependent variable can be explained by an independent variable.

The formula for R2 score is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where:

R^2 is coefficient of determination

n is number of observations,

y_i is actual value of the dependent variable for observation i ,

\hat{y}_i is predicted value of the dependent variable for observation i ,

\bar{y} is mean of the actual values of dependent variable.

An R2 score of 1 indicates a perfect fit of 100% accuracy, while a score of 0 indicates that the model does not explain any variability in the dependent variable.

4.2 Mean Squared Error (MSE)

The Mean Squared Error (MSE) is a measure of the average squared difference between predicted and actual values. It quantifies the average of the squares of the errors.

The formula for MSE is given by:

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

where:

MSE is the mean squared error,

n is the number of observations,

y_i is the actual value of the dependent variable for observation i ,

\hat{y}_i is the predicted value of the dependent variable for observation i .

A lower MSE indicates better model performance as it signifies smaller errors between predicted and actual values.

4.3 Mean Absolute Error(MAE)

Mean Absolute Error (MAE) is one of the most accurate criterions often used to evaluate the exactitude of a forecasting or predication model. Absolute difference between the predicted and real values is the average value of the difference that it equals. MAE offers a simple measure for the severity of mistakes in a group of predictions, taking into account the scale regardless of the direction these mistakes might have.

The Mean Absolute Error (MAE) is calculated as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where:

n = Number of observations

y_i = Actual value

\hat{y}_i = Predicted value

4.4 Top-M

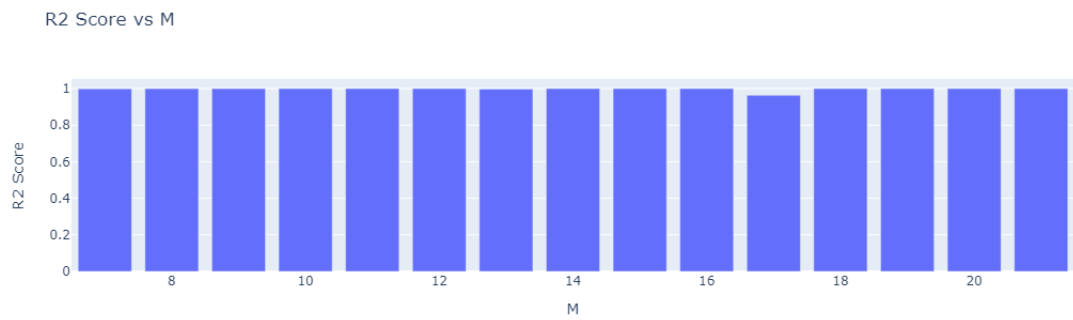


Figure 4.1: R2 score for M Cluster

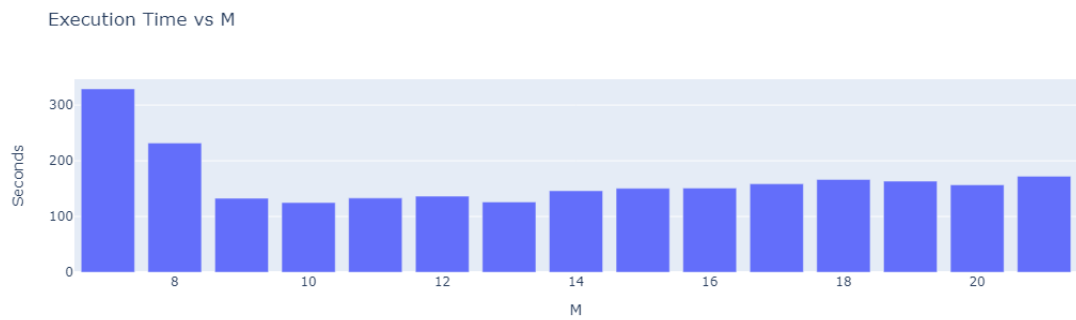


Figure 4.2: Execution Time For Top M Cluster

4.5 Cluster

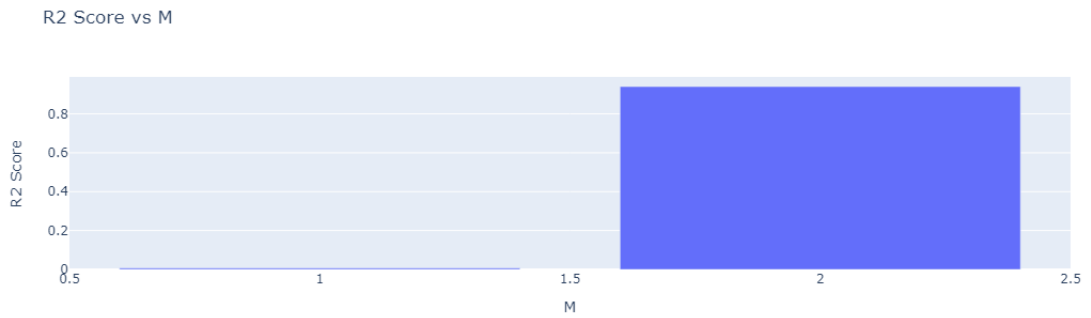


Figure 4.3: R2 score for Top M Feature lies between Median Range

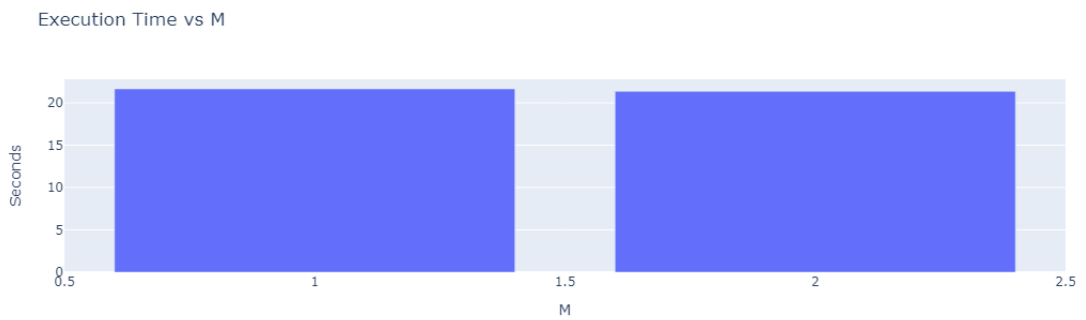


Figure 4.4: Execution Time For Top M feature Lies Between Median Range

4.6 Median Range

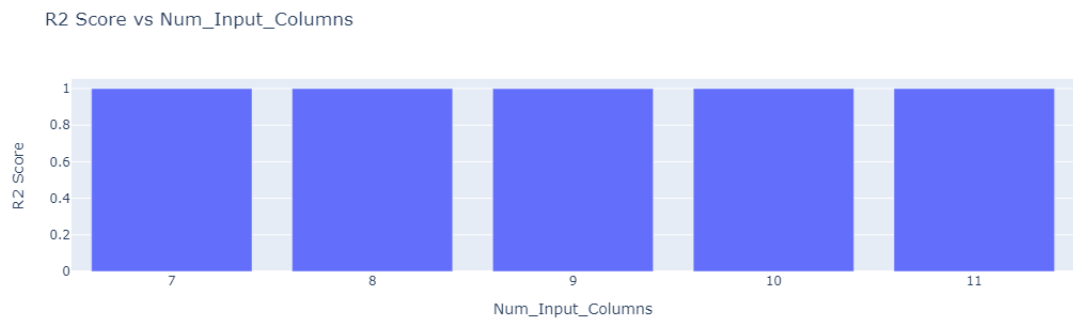


Figure 4.5: R2 score for Top M Feature

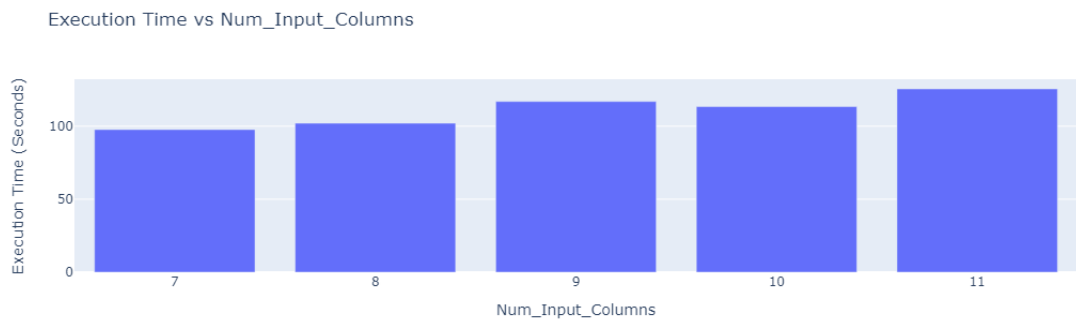


Figure 4.6: Execution Time For Top M feature

4.7 Independent Components Analysis(ICA)

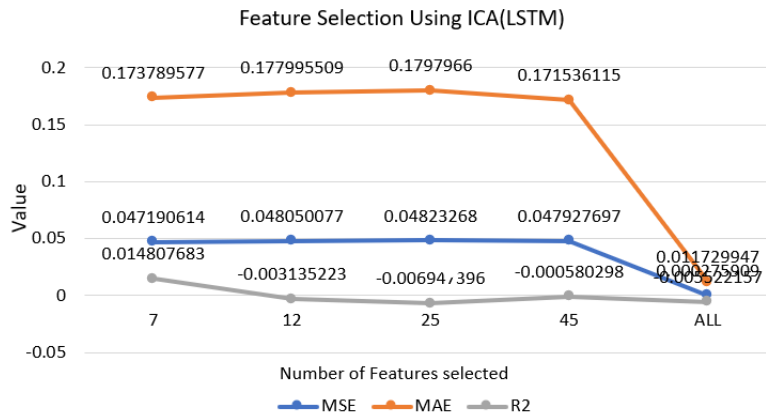


Figure 4.7: accuracy comparison for 48 Features using ICA (LSTM)

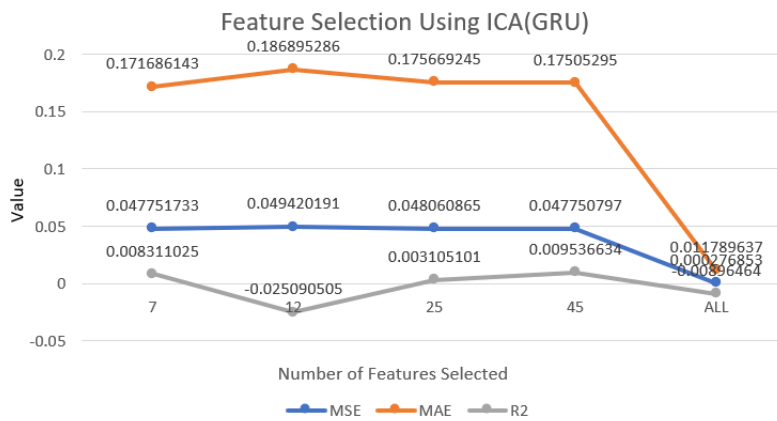


Figure 4.8: accuracy comparison for 48 Features using ICA (GRU)

4.8 GA-LSTM

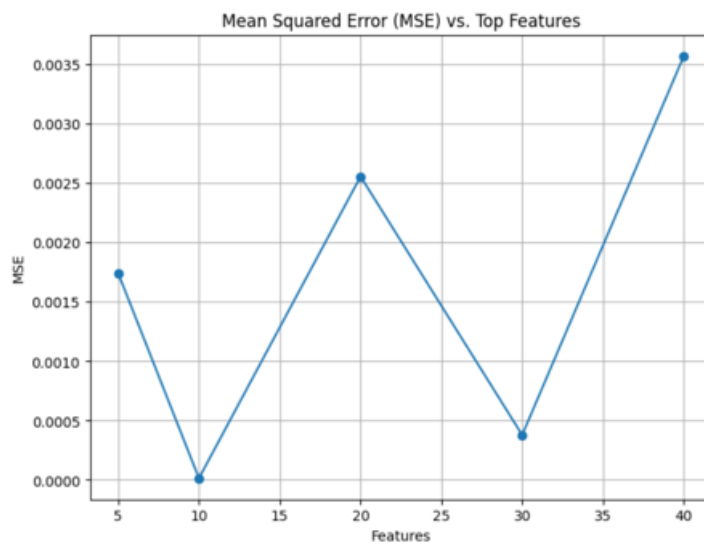


Figure 4.9: MSE comparison for 40 Features using GA-LSTM(CCB)

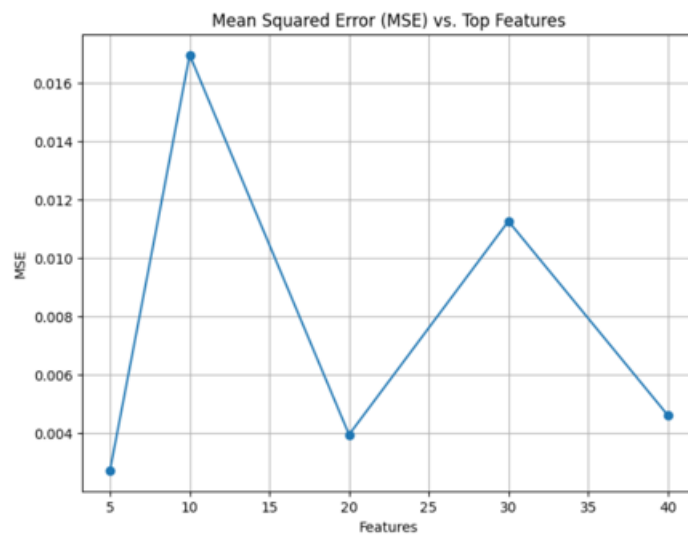


Figure 4.10: MSE comparison for 40 Features using GA-LSTM(CSI 300)

Chapter 5

Conclusion remarks and future directives

5.1 Summary

The “Stock Trend Prediction” project aims to harness the predictive power of Long Short-Term Memory network, CNN and GRU to analyze and forecast stock price movements. The methodology follows a sequence of collecting historical stock data, pre-processing it to address inconsistencies and prepare it for analysis, and constructing a neural network architecture specialized in handling temporal dependencies. The Deep learning model is meticulously trained on the processed data to minimize forecast errors, with particular attention to preventing overfitting and optimizing hyperparameters for peak performance. Upon validation and testing, the model’s predictive accuracy is evaluated using appropriate metrics, and the outcomes are visualized to facilitate interpretation of the results. Should the model prove robust and reliable, it could be deployed to provide real-time insights, thus serving as an innovative tool for investors aiming to navigate the complex dynamics of the stock market with data-driven strategies.

5.2 Conclusion

Identifying non-linear stock market trends is tough. Stock market data may be studied using various computational approaches, with characteristics playing a key role in their effectiveness. They analyse difficult data and provide valuable insights. Therefore, identifying key traits is crucial for accurate market predictions.

In this project we used CV based selection method to forecast the results. We also use ICA and Genetic algorithm to predict the movement of the stock. And in CV based method we are also using 3 sub method like Top M where we are selecting feature top feature by CV values. Then we are using K-means algorithm to group the feature and then we feed to the deep learning method and in last we are using median range to exclude feature are outliers in the Median range of the CV values. From the results we can see that using CNN model we are getting more accurate results then other method.

5.3 Future Directive

As we know Stock market is very broad topic and very difficult to predict. And it is also not dependent only on some things like previous price or calculated features or etc. Stock price is also dependent on fundamental of their company, overall market condition, in which segment that stock belongs to, Region, their current results of performance and so many things. More ever what news there is in certain market regarding that stock also effect the price so we can create that model which calculate the impact of that news that affect the stock price. And that should be update on regular bases both stock price and news which are related to overall market or for that particular stock automatically. Also we can include some fundamental data for that stock to predict more accurately.

Bibliography

- [1] K. Chaudhari and A. Thakkar, "Neural network systems with an integrated coefficient of variation-based feature selection for stock price and trend prediction," *Expert Systems with Applications*, vol. 219, p. 119527, 2023.
- [2] G. Mu, N. Gao, Y. Wang, and L. Dai, "A stock price prediction model based on investor sentiment and optimized deep learning," *IEEE Access*, 2023.
- [3] L. Wang, Q. Cao, Z. Zhang, S. Mirjalili, and W. Zhao, "Artificial rabbits optimization: A new bio-inspired meta-heuristic algorithm for solving engineering optimization problems," *Engineering Applications of Artificial Intelligence*, vol. 114, p. 105082, 2022.
- [4] Y. Zhao and G. Yang, "Deep learning-based integrated framework for stock price movement prediction," *Applied Soft Computing*, vol. 133, p. 109921, 2023.
- [5] J. Chen, Y. Wen, Y. Nanekaran, M. Suzauddola, W. Chen, and D. Zhang, "Machine learning techniques for stock price prediction and graphic signal recognition," *Engineering Applications of Artificial Intelligence*, vol. 121, p. 106038, 2023.
- [6] B. M. Henrique, V. A. Sobreiro, and H. Kimura, "Practical machine learning: Forecasting daily financial markets directions," *Expert Systems with Applications*, p. 120840, 2023.
- [7] L. Cagliero, J. Fior, and P. Garza, "Shortlisting machine learning-based stock trading recommendations using candlestick pattern recognition," *Expert Systems with Applications*, vol. 216, p. 119493, 2023.

- [8] B. Gülmez, “Stock price prediction with optimized deep lstm network with artificial rabbits optimization algorithm,” *Expert Systems with Applications*, vol. 227, p. 120346, 2023.
- [9] H. Song and H. Choi, “Forecasting stock market indices using the recurrent neural network based hybrid models: Cnn-lstm, gru-cnn, and ensemble models,” *Applied Sciences*, vol. 13, no. 7, p. 4644, 2023.
- [10] K. Prachyachuwong and P. Vateekul, “Stock trend prediction using deep learning approach on technical indicator and industrial specific information,” *Information*, vol. 12, no. 6, p. 250, 2021.
- [11] A. Sethia and P. Raut, “Application of lstm, gru and ica for stock price prediction,” in *Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 2*, pp. 479–487, Springer, 2019.
- [12] S. Chen and C. Zhou, “Stock prediction based on genetic algorithm feature selection and long short-term memory neural network,” *IEEE Access*, vol. 9, pp. 9066–9072, 2020.
- [13] “Financial Terms Dictionary — investopedia.com.” <https://www.investopedia.com/terms>.
- [14] “Coefficient of variation - Wikipedia — en.wikipedia.org.” https://en.wikipedia.org/wiki/Coefficient_of_variation.
- [15] H. B. Zhou and J. T. Gao, “Automatic method for determining cluster number based on silhouette coefficient,” *Advanced materials research*, vol. 951, pp. 227–230, 2014.

22MCED10_Major_Part_2_removed

ORIGINALITY REPORT

7%

SIMILARITY INDEX

4%

INTERNET SOURCES

6%

PUBLICATIONS

3%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Greenwich Student Paper	1%
2	www.mdpi.com Internet Source	1%
3	fastercapital.com Internet Source	1%
4	Yuancheng Si, Saralees Nadarajah, Zongxin Zhang, Chunmin Xu. "Modeling opening price spread of Shanghai Composite Index based on ARIMA-GRU/LSTM hybrid model", PLOS ONE, 2024 Publication	1%
5	Kinjal Chaudhari, Ankit Thakkar. "Neural network systems with an integrated coefficient of variation-based feature selection for stock price and trend prediction", Expert Systems with Applications, 2023 Publication	1%
6	Submitted to The Robert Gordon University Student Paper	1%

7

Submitted to University of St Andrews

Student Paper

1 %

8

www.tutorialspoint.com

Internet Source

1 %

Exclude quotes On

Exclude matches < 1%

Exclude bibliography On