ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Forecasting Ethiopian Agricultural Commodity Price Using Time Series Features and Technical Indicators

---

By:
Sisay Gebremedhin

Advisor:
Dr. Surafel Lemma

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Engineering

May 17, 2022
Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

The undersigned have examined the thesis titled:

Forecasting Ethiopian Agricultural Commodity Price Using Time Series Features and Technical Indicators

Presented by Sisay Gebremedhin, a candidate for the degree of Master of Science and hereby certify that it is worthy of acceptance.

Approved By Board of Examiners

Dr. Bisrat Derebssa	_____
Dean, SECE, AAiT	Signature
Dr. Surafel Lemma	_____
Thesis Advisor	Signature
Dr. Fitsum Assamenew	_____
Examiner I	Signature
Dr. Sosina M. Gashaw	_____
Examiner II	Signature

Declaration of Authorship

I, Sisay Gebremedhin, declare that this thesis titled, "Forecasting Ethiopian Agricultural Commodity Price Using Time Series Features and Technical Indicators" and the work presented in it are my own. I confirm that:

- This work was completed entirely or primarily while in candidature for a research degree at this university.
- This has been clearly stated if any part of this thesis has previously been submitted for a degree or other qualification at this University or any other institution.
- Where I have referenced the published work of others, this is always carefully acknowledged.
- I always give the source when I quote from other people's work. Except for such quotes, this thesis is entirely my work.
- I have acknowledged all main sources of help.
- Where the thesis is based on a cooperative effort by myself and others, I have specified clearly what was done by others and what I contributed.

Signed:

_____

Date:

_____

Abstract

Agricultural commodity price prediction helps the government, investors, and farmers to make informed decisions. Realizing the benefit, several researchers proposed different prediction models that use different features. However, most prediction models are affected by factors, such as data type (e.g., linear and nonlinear), seasonality of commodity items, weather conditions, commodity volatility features, and country economic factors. Among these factors, the most significant impediments to the accuracy of commodity price prediction are seasonality and trend pattern. To fill this gap, we propose a model that predicts commodity prices through the combination of time series features and technical indicators. The prediction model is built using four-machine learning algorithms: Artificial Neural Network, Extreme Learning Machine, Support Vector Machine, and Random Forest. To assess the impact of the proposed approach, we conducted two experiments using coffee and sesame datasets. The performance of the prediction models is assessed using the root mean square error (RMSE) and mean average error (MAE). The results show that the proposed approach improves agricultural commodity price prediction performance in all cases except MAE of sesame while using Extreme Learning Machine. Using Artificial Neural Network, Extreme Learning Machine, Support Vector Machine and Random Forest, the RMSE of price prediction is reduced by an average of 4.37, 4.42, 2.74, and 5.15, respectively. Finally, among the four machine learning algorithms used in the study, Artificial Neural Network is found to be the best algorithm for enhancing the performance of agricultural commodity price prediction. We also conclude from our experiment result that considering commodity properties such as periodicity, volatility, linearity, momentum, volume, and trend would improve the performance of agricultural commodity price prediction. To see which of the features contributed more to the improvement of agricultural commodity price prediction, we computed feature importance using Random forest algorithms. The result shows that: close, high, low, open, exponential moving average (EMA), double exponential moving average (DEMA), simple moving average (SMA), truehigh, truelow, trend, seasonality, relative strength index (RSI) are the most important features in sesame and coffee price prediction.

Keywords : *technical indicator, time series feature, price forecasting, agricultural commodity.*

Acknowledgements

In the first place, I want to forward my sincere thanks to God for standing at my side regardless of different challenges and ups and downs. Next, I want to thank my advisor, Dr. Surafel Lemma, for his consistent guidance and supervision. I would also want to thank the ECX personnel for their full cooperation in providing data for my research. I would also like to express my heartfelt gratitude to my family for their valuable assistance in all aspects of my life.

Contents

Declaration of Authorship i

Abstract ii

Acknowledgements iii

1 Introduction 1
- 1.1 Problem Statement . . . 2
  - 1.1.1 Research Question . . . 3
- 1.2 Objective . . . 3
  - 1.2.1 General Objective . . . 3
  - 1.2.2 Specific Objective . . . 3
- 1.3 Research Methodology . . . 4
- 1.4 Scope . . . 4
- 1.5 Significance of the Study . . . 4
- 1.6 Contribution . . . 5
- 1.7 Paper Organization . . . 5

2 Theoretical Background 6
- 2.1 Time Series Analysis . . . 6
  - 2.1.1 Seasonal Component . . . 6
  - 2.1.2 Trend Component . . . 6
- 2.2 Technical Analysis . . . 7
  - 2.2.1 Trend Indicators . . . 8
    - Moving Averages . . . 8
  - 2.2.2 Momentum Indicator . . . 10
  - 2.2.3 Volatility Indicator . . . 12
  - 2.2.4 Volume Indicator . . . 12
- 2.3 Different Machine Learning Algorithms . . . 13
  - 2.3.1 Extreme Learning Machine (ELM) . . . 13
  - 2.3.2 Random Forest Algorithm . . . 15
  - 2.3.3 Artificial Neural Networks . . . 16
  - 2.3.4 Support Vector machine . . . 16
- 2.4 Feature selection Algorithm . . . 18

2.4.1 Correlation-based Feature Selection (CFS) . . . 18

3 Literature Review **20**
3.1 Agricultural Commodity Price Prediction Using Traditional Method . . . 21
3.1.1 Regression Analysis Forecasting Method . . . 21
3.1.2 Time Series Analysis Forecasting Method . . . 21
3.2 Agricultural Commodity Price Prediction Using Intelligence Method . . . 22
3.2.1 Artificial Neural Network (ANN) . . . 23
3.2.2 Support Vector Machine (SVM) . . . 24
3.3 Hybrid Forecasting Method . . . 24
3.3.1 General Hybrid Model . . . 25
3.3.2 Decomposition-Based Hybrid Model . . . 25

4 Proposed Approach **27**
4.1 Data Collection . . . 27
4.2 Data Preprocessing . . . 29
4.2.1 Handling Outliers . . . 29
4.2.2 Data Cleaning . . . 29
4.3 Feature Extraction . . . 30
4.3.1 Technical Indicator . . . 30
4.3.2 Time Series Features . . . 30
4.4 Feature Reduction . . . 32
4.5 Model Building . . . 32
4.5.1 Artificial Neural Network (ANN) . . . 32
4.5.2 Support Vector Machine (SVM) . . . 32
4.5.3 Random Forest Algorithm (RF) . . . 33
4.5.4 Extreme Learning Machine (ELM) . . . 33
4.6 Prediction and Evaluation . . . 33

5 Experiments **34**
5.1 Dataset Description . . . 34
5.1.1 Feature Description and Analysis . . . 34
5.2 Experiment Setup . . . 35
5.2.1 Extreme Learning Machine Model Setup . . . 36
5.2.2 Support Vector Machine Model Setup . . . 36
5.2.3 Artificial Neural Network Model Setup . . . 37
5.2.4 Random Forest Model Setup . . . 37
5.2.5 Experiment Environment . . . 38
5.3 Evaluation Metrics . . . 38

5.3.1 Mean Absolute Error (MAE) . . . 38
5.3.2 Root Mean Squared Error (RMSE) . . . 38
5.3.3 Root Relative Squired Error (RRSE) . . . 39
5.3.4 Relative Absolute Error (RAE) . . . 39
5.4 Results . . . 40
5.4.1 Experiment I . . . 40
5.4.2 Experiment II . . . 44
5.5 Discussion . . . 46
5.6 Threats to Validity . . . 47
5.6.1 Threats to Internal Validity . . . 47
5.6.2 Threats to External Validity . . . 47

6 Conclusion And Recommendation **48**
6.1 Conclusion . . . 48
6.2 Recommendation and Future works . . . 49

Bibliography **50**

A Feature Description and Analysis **54**

List of Figures

2.1 The working principle of Extreme Learning Machine [19]. 14
2.2 The working principle of Random Forest algorithm [20]. 15
2.3 The working principle of Artificial Neural Network [21]. 16
2.4 The working principle of Support Vector Machine [14]. 17

3.1 Literature Review Framework [24]. 20

4.1 Proposed Approach. 28

5.1 Sesame price prediction using Artificial Neural Network. 43
5.2 Sesame price prediction using Support Vector Machine. 43
5.3 Coffee price prediction using Artificial Neural Network. 43
5.4 Coffee price prediction using Support Vector Machine. 44
5.5 Sesame feature importance chart using Random Forest. 45
5.6 Coffee feature importance chart using Random Forest. 45

A.1 Feature correlation matrix. 55
A.2 correlation matrix After feature reduction. 56

List of Tables

4.1 Features used in this study. 31

5.1 Summary of dataset. 34
5.2 The final features after feature reduction. 35
5.3 Machine Specification. 38
5.4 Result using Artificial Neural Network. 41
5.5 Result using Extreme Learning Machine. 41
5.6 Result using Support Vector Machine. 41
5.7 Result using Random Forest. 42
5.8 Comparative analysis of the applied algorithms. 42

A.1 The reserved features after feature reduction. 54

List of Abbreviations

ADL	Accumulation Distribution Line
ADX	Average Directional Movement Index
AE	Absolute error
AI	Artificial Intelligence
ANN	Artificial Neural Network
ARIMA	Auto Regressive Integrated Moving Average
ATR	Average True Range
ATTLSTM	Attention Based Long-Short Term Memory
BF	Best First
BNN	Bayesian Neural Network
BPNN	Back-Propagation Neural Network
CC	Correlation Coefficient
CCI	Commodity Chanel Index
CFS	Correlation Based Feature Selection
CMF	Chaikin Money Flow
CMO	Chande's Momentum Oscillator
CNN	Convolutional Neural Network
DEMA	Double Exponential Moving Average
ECX	Ethiopian commodity Exchange
ELM	Extreme Learning Machine
EMA	Exponential moving Average
EMD	Empirical Mode Decomposition
EPS	Earning Per-Share
ES	Exponential Smoothing
GA	Genetic Algorithm
GDGA	Gradient Decent Genetic Algorithm
HGWO	Hybrid Grey Wolf Optimization
IG	Information Gain
IPSO	mproved Particle Swarm Optimization
KNN	K-Nearest Neighbor
MACD	Moving Average Convergence/Divergence
MAE	Mean Absolute Error
MEA	Means End Analysis
MFI	Money Flow Index
OBV	On Balance Volume
PSO	Particle Swarm Optimization
QR	Quantile Regression
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Artificial Neural Network
RSI	Relative Strength Index

SARIMA	Seasonal Auto Regressive Integrated Moving Average
SIA	Seasonal index adjustment
SLFN	Single Hidden layer Feedforward Neural Network
SMA	Simple Moving Average
SMI	Stochastic Momentum Index
SVC	Support Vector Classification
SVM	Support Vector Machine
SVR	Support Vector Regression
STL	Seasonal and Trend Losses
TDI	Trend Detection Index
TP	Typical Price
TS	Time Series
VAR	Vector Auto Regression
VHF	Vertical Horizontal Filter
VMD	Variable Mode Decomposition
WA	Wavelet Analysis
WPR	WilliamR– Williams Percentage Range
WPT	Wavelet Packet Transform

Dedicated to my families (G/medhin Abraha, Letebirhane Beyene, Alemseged, Negasi, and Mahlet)

Chapter 1

Introduction

Artificial intelligence (AI) is a new concept that has been used in the development of autonomous automobiles, intelligent robotics, image and speech recognition, automatic translations, and medical assistance. In light of AI and the application of different machine-learning algorithms, forecasting agricultural commodities has become crucial in the financial and economic fields [1]. Consequently, over the last ten years, scholars have considered developing an accurate prediction model. The aim to predict commodity prices more correctly drives researchers to work on improving the present predictive machine learning models. Stockholders have the freedom to establish plans and strategic approaches to investments and future activities. Consequently, groups and individuals seek any prediction strategy that promises increased income from the commodity market with minimal risk [1].

Forecasting commodity prices is among the most challenging issues in finance due to the stochastic behavior and complex interdependence of the agricultural commodity market [2]. The noisy(rough) and non-stationary nature of agricultural commodities futures makes prediction difficult [2]. The term ”noisy” refers to the insufficient information available to observe past agricultural commodity futures behavior. The term ”non-stationary” refers to agricultural commodities futures can move substantially over time. These characteristics contribute to poor agricultural commodities futures prediction outcomes when classic empirical models like as linear models, Vector Auto-Regression, and Auto-Regressive Integrated Moving Average [3]–[5]. The above methods are work premised on the idea that variables are independent and have a normal distribution, which contradicts real-world market conditions.

In recent years, machine learning methods have been applied in time series forecasting [6]. Example of such models are Support Vector Regression, Artificial Neural Networks, and Bayesian Neural Network [7]. In addition, hybrid approaches were developed to boost prediction efficiency. However, there is limited proof of their relative performance as conventional prediction models. In particular, for short-term prediction, ARIMA is extensively employed for its efficiency in time series prediction than most other neural network algorithms [8].

To predict commodity prices, approach such as fundamental analysis,time-series analysis and technical analysis, are used. The fundamental analysis focus on the driving factors of pricing, such as terrestrial and climate conditions, artificial or natural disasters, and political data, for predicting future prices [9], [10]. Technical analysis is the assessment of securities/assets based on market activity statistics such as previous prices and volume [10], [11]. As a result, technical analysis is also known as behavioral finance and is the study of human mass psychology. Price chart analysts look for price patterns and use price data in different technical indicator assessments to predict future price direction. Time series analysis is a stochastic approach for studying and modifying time series data. A time series is a collection of data points gathered at periodic intervals [12]. In contrast to regression analysis, time series analysis is very helpful for locating the primary features and statistics of time series data. Time series analysis offers aspects that can help us understand the underlying variables that lead to a given pattern in time series data points and, as a result, can help us anticipate data points.

In this research, we propose an approach for forecasting agricultural commodity prices that takes into account both technical indicators and time series features. This system extracts twenty eight features from agricultural commodity prices based on seasonality, trend, volatility, linearity momentum, and volume. As candidate prediction models, intelligent forecast models such as Artificial Neural Network, Support Vector Machine, Extreme Learning Machine, and Random Forest are selected. The regressor identifies the relation between extracted feature and the model's accuracy.

The result shows that the proposed approach improves the agricultural commodity price prediction performance in 7 out of 8 cases. The RMSE of price prediction is reduced by an average of 4.37, 4.42, 2.74, and 5.15 when using Artificial Neural Network, Extreme Learning Machine, Support Vector Machine, and Random Forest, respectively. Besides, we also compare the importance level of each feature for both the sesame and coffee commodity by computing the feature importance using Random Forest algorithm.

1.1 Problem Statement

Agricultural commodity price prediction helps the government, investors, and farmers to make informed decisions. Realizing the benefit, several researchers proposed different prediction models that use different methods [10][13][14][15]. However, most prediction models are affected by factors, such as data type (e.g., linear and nonlinear), seasonality of commodity items, weather conditions, commodity volatility features, and country economic factors. Among these factors, the most significant impediments to the accuracy of commodity price prediction are seasonality and trend pattern. Research in Ethiopia's

agricultural commodity price forecasting is vital to reduce the risks of fluctuating prices. The previous study conducted considering in Ethiopia context [13][16] has two main shortcomings. First, when forecasting agricultural commodity prices, periodicity properties like seasonality and trend patterns were not considered. Second, while predicting commodity prices, time-series features of agricultural commodities such as volatility, volume, momentum, trend, and linearity were not considered.

Therefore, the objective of this research is to examine the impact of periodicity and other time series characteristics on Ethiopian agricultural commodity price forecasting. We will study the effect and capability of seasonal and trend patterns in Ethiopian agricultural commodity prices prediction. We also aim to predict agricultural commodity future prices simultaneously with full consideration of time series characteristics such as momentum, volatility, trend, volume, and linearity. To this end, we will address the following research question.

1.1.1 Research Question

1. Does considering time series seasonal and trend pattern (periodicity) properties of the commodities improve price-forecasting performance?
2. With regard to agricultural commodity price prediction scheme, what are the important features that help give better forecasting?

1.2 Objective

1.2.1 General Objective

The general objective of this research is to examine the impact of seasonality and trend pattern (periodicity), momentum, volatility, trend, volume, and linearity properties of agricultural commodity towards price prediction.

1.2.2 Specific Objective

- To assess the impact of seasonality and trend pattern (periodicity) in commodity price prediction
- To assess the dynamic dependencies among different time series features.
- To compare performance of machine learning algorithms in price prediction system.
- To examine the importance of feature in the price prediction approach.

1.3 Research Methodology

While conducting this research, steps outlined below are followed:

- **Literature Review**: Throughout the research, literature have been reviewed to formulate the problem, identify gaps in prior works and propose an approach.
- **Propose an Approach**: Once a clear problem is formulated, an approach that can be a solution for the stated problem is proposed.
- **Data Collection and Preparation**: coffee and sesame data are collected, and from the collected data, features are generated by applying required feature extraction methods.
- **Verify Proposed Approach**: Finally, the proposed approach is evaluated using prepared test data.
- **Results and Evaluation**: In order to measure the performance of our approach, we used commonly used evaluation metrics like RMSE and MAE.

1.4 Scope

This research focuses on time series forecasting for Ethiopian agricultural commodities' future market prices. The scope of this study is limited to two Ethiopian agricultural products: coffee and sesame. However, this study cannot consider irregular components of price movement such as pandemics, artificial or natural disasters, climate conditions, and political situations. In general, the following core points can summarize the scope of this thesis:

- Preparing the dataset for the proposed time series commodity price prediction
- Examining whether the identified price prediction features have an impact on commodity price prediction.
- Examining the capability of the proposed approach on the existing commodity price prediction scheme.

1.5 Significance of the Study

This study aims to avail in real-time, integrated, and up-to-date prediction of agricultural commodities (like coffee and sesame) price information to farmers, traders, and decision-makers. Furthermore, it will provide a brief indication for Ethiopian policymakers and

strategists to prepare effective policies based on the expected price.

1.6 Contribution

This research work aimed to assess the impact of seasonality and trend pattern (periodicity) and other time series characteristics such as volatility, volume, momentum, and linearity in agricultural commodity price series. Therefore, the main contributions of this research work are as follows:

- Prepared a time series dataset for agricultural commodity price prediction.
- Assessed the impact of seasonality and trend pattern (periodicity) and other time series properties such as volume, stationarity, volatility, and momentum in agricultural commodity price prediction.
- For training and prediction purpose the features used in the prediction methodology is the combination of time series features and technical indicators. This makes it unique from other works and also improves the prediction accuracy.
- Identified the important features for agricultural commodity price prediction.

1.7 Paper Organization

The thesis document consists of six chapters: The first chapter deals with the motivation, the general and specific objectives, the scope, and the contribution of the thesis. The next chapter presents the details of theoretical backgrounds of different machine learning algorithms, technical analysis and time series analysis. The third chapter makes a detailed investigation of previously done related works of literature. The chapter four describes the proposed methodology. The fifth chapter presents the experiment setups of the entire thesis which includes: the algorithms hyperparameter configuration, the evaluation metrics used, and the validation techniques and gives a detailed explanation of the results obtained from the conducted sets of experiments. The final chapter concludes the thesis and indicates future research directions.

Chapter 2

Theoretical Background

This chapter discusses the theoretical background of some core components of this research work. It briefly explain the working principles, background, and main components that build up machine-learning algorithms. In addition, this section is dedicated to give a brief introduction about market price analysis such as technical analysis and time series analysis.

2.1 Time Series Analysis

Time series analysis is a statistical method for analyzing and manipulating data arranged in a time-series manner. Time-series data is recorded at a specified time interval. The interval between the recorded data is equal. In contrast to regression analysis, time series analysis is particularly useful for obtaining meaningful statistics and properties of a time series dataset. The time series analysis can detect the underlying variables that lead to a specific pattern in the data series points. The first step in time series analysis is to determine whether the data set is stationary or nonstationary. Time series is considered as stationary when statistical properties, such as mean, variance, and autocorrelation, are constant over time. Otherwise, the time series data considered as non stationary. The main reasons for non-stationarity are seasonality and trend components of the time series data [17], [18].

2.1.1 Seasonal Component

The seasonal component is referred to as seasonality of a time series data. The seasonal variable reflects “normal” variations that occur every year to the same extent. Example of such variation are, whether fluctuations, and cultivation time of commodity. Seasonality has a fixed and predictable periodicity. For example, monthly coffee and sesame sales show seasonality due to changes in commodity costs at the end of the calendar year [18].

2.1.2 Trend Component

The long-term pattern of time series data is represented by the trend. The trend could be positive or negative depending on whether the time-series pattern is up or down. When

there is no fluctuating pattern in a time series, the data is said to be stationary [18]. The following tests will be conducted to check whether or not a time series is stationary;

- Plotting Rolling Statistics: this testing mechanism Plots the moving average or difference and checks whether it depends on time (time varies). It is more like a visual representation.
- Dickey-Fuller Test is a statistical test used to determine the stationarity of time series, unlike the first one. This test assumes the null hypothesis is not stationary to the time series. The test-statistic and some critical values for various confidence levels are the results. When the test-statistical falls below the critical value, the null hypothesis is rejected, the series is stationary.

The purpose is to eliminate these characteristics from the time series data by evaluating its trend and seasonality. The following techniques can be used for modeling or estimating trend and seasonality:

- Smoothing: Considers rolling averages.
- Aggregation: Considers averages for time period like month/week.
- Polynomial Fitting: Fits a regression model

Most of these methods are used in various problem-solving procedures. After estimation, the following strategies can eliminate trends and seasonality:

- Differencing: is the most common strategy to reduce non-stationary features by calculating the difference between a current observation and the previous one. The sequence of differences can be modified to optimize trend and seasonal reduction.
- Decomposing: is the tool for all kinds of time series analysis, especially seasonal adjustment. It is intended to build a component that could be employed by adding or multiplying the original data in an observation series. A distinct sort of behavior characterizes each of the features.

2.2 Technical Analysis

The technical analysis calculate statistical trends like price and volume changes in stocks which are intended to spoil the opportunities. The presumption is that the known foundations are price, and hence they must be carefully addressed. Technical analysts are not trying to estimate a security's inherent value. Comparatively, they use inventory charts to detect patterns and trends that show how an inventory will function in the future.

This section presented a details of all technical indicators, which are divided into four categories: trend, volume, momentum, and volatility indicators [10], [11].

2.2.1 Trend Indicators

Trend indicators are used to measure the trend's direction and strength with a price average in some form to determine the basis. When pricing is higher than usual, it indicates buy. When it moves below average, it indicates a sell signal [11].

Moving Averages

Moving average is a kind of pulse response filter used to analyze particular data points by building an average set of sub-sets of the entire data set. These methods are employed for noise reduction via standardization of the data and trends. It is often used for mixing many moving averages due to its simplicity and versatility [11].

Simple moving average (SMA): assigns the same amount of weight for each day. It is the biggest drawback of the simple moving average [10]. Equation 2.1 shows the mathematical formulation of the SMA.

$$SMA = (Pc(t) + Pc(t-1) + \cdots + Pc(t-n-1))/n \tag{2.1}$$

Exponential moving average (EMA): unlike SMA, the exponential moving average gives higher priority to the actual data. In comparison to the furthest values, the current values become more important in the EMA computation [10]. EMA is calculated using Equation 2.2.

$$\text{EMA} = \text{EMA}_{(-1)} + \text{Kx}\left(\text{ input } - \text{EMA}_{(-1)}\right) \text{ where K} = \frac{2}{(\text{n}+1)} \tag{2.2}$$

Double Exponential Moving Average (DEMA): is closer to the price history than most other moving averages so that the delay is lower and the curve is not so choppy [10]. DEMA is calculated using Equation 2.3.

$$\text{DEMA} = 2 \times \text{EMA}(n) - \text{EMA}(EMA(n)) \tag{2.3}$$

In all moving averages, when close prices are above the moving average, they indicate a buy signal. In addition, when it is below its moving average, it signals sell.

Moving Average Convergence/Divergence (MACD) Indicator: MACD is a momentum oscillator that shows the dynamics and intensity of the current trend and moves in both directions around the zero line. There are three moving averages of MACD. They usually have 9, 12, and 26 settings [11]. Finally there is the histogram, which is the difference between the MACD and the signal as shown in Equation 2.4.

$$\begin{aligned} MACDline &= 12dayEMA - 26dayEMA \\ Signalline &= 9dayEMAofMACDline \\ EMA &= ExponentialMovingAverage \end{aligned} \tag{2.4}$$

Average Directional Movement Index (ADX): ADX assesses the strength or weakness of a trend instead of the true direction. The directional movement is defined as +DI and -DI. By indicating the trend direction, the Plus Indicator, +DI and the Minus Indicator (-DI) complement ADX. The +DI is the proportion of the true range. The DI is the proportion of the true range that is down. The values range between 0 and 100 but rarely exceed 60. The high ADX number shows a strong trend and a low one shows a weak trend. Equation 2.5 shows the mathematical formulation of the ADX [11].

$$\mathrm{ADX} = \frac{\mathrm{ADX}_{(-1)} \times (\mathrm{n}-1) + \mathrm{DX}}{\mathrm{n}} \mathrm{DX} = \frac{(+\mathrm{DI}) - (-\mathrm{DI})}{(+\mathrm{DI}) + (-\mathrm{DI})} \tag{2.5}$$

Trend Detection Index (TDI): is used to determine the start and end of a trend. The Trend Detection Index employed as an independent indicator or in combination with others; it works well when trend starts are recognized. A positive indicator value indicates an upward trend, although a negative number shows a downward trend [11]. Equation 2.6 determine the start and end of a trend.

$$\text{20 - day Trend Detection Index (TDI)} = (\ \mathrm{AV20}\ ) - \{(\Sigma\ \mathrm{AM}\ 40) - (\Sigma\ \mathrm{AM}\ 20)\} \tag{2.6}$$

Aroon Indicator (Aroon): tries to demonstrate the dawn of a new trend. The indicator is made up of two axes (up and down) that measure how long it has remained since the highest / lowest was in a period of n. When the Aroon Up remains between 70 and 100, the trend is up. When Aroon Down stays between 70 and 100, it shows a downward trend [11]. Equation 2.7 shows the mathematical formulation of the Aroon up and aroon down.

$$\begin{aligned} \mathrm{AroonUP} &= 100 \times \left(\frac{\mathrm{n} - \ \mathrm{PeriodSinceHighestHigh}}{\mathrm{n}} \right) \\ \mathrm{AroonUP} &= 100 \times \left(\frac{\mathrm{n} - \ \mathrm{PeriodSinceLowestLow}}{\mathrm{n}} \right) \end{aligned} \tag{2.7}$$

Vertical Horizontal Filter (VHF): determines the trend in prices. The increase of VHF shows the development of a trend. Increased VHF values show a stronger trend. When the VHF is down, the trend ends and the price is congested [11]. VHF is calculated using Equation 2.8.

$$\text{VHF} = \frac{\text{HighestHigh} - \text{LowestLow}}{\sum_{1}^{\text{n}} \frac{\text{close i} - \text{close}}{\text{i}-1}} \text{close i} - 1 \tag{2.8}$$

2.2.2 Momentum Indicator

Momentum indicators assist traders to identify price movement through price comparison over time. This type of indicator is usually utilized for price analysis and volume analysis [11].

Relative Strength Index (RSI): has been employed in the financial industry as a technical indicator. This seeks to map current and past market strengths or weaknesses based on closing pricing in the latest trade session. The RSI varies from 0 to 100. If the RSI is above 70, it signals overbought and below 30 shows oversold. If the prices increase and the RSI does not, the reversal will be indicated [11]. RSI is used for calculating the speed and pattern of price fluctuations, which is given by Equation 2.9.

$$RSI = 100 - \left(\frac{100}{1 + \frac{U}{D}} \right) \tag{2.9}$$

Where

U- average value of the positive price changes over n days;
D- average value of the negative price changes over n days.

Stochastic Oscillator (Stoch): assesses the relationship between the close and the most recent trading range. The scale runs from 0 to 100. Values greater than 75 indicate an overbought condition, while values less than 25 suggest an oversold position. When the Fast percent (D) crosses above the Slow percent (D), it signals a buy; when it crosses below, it signals a sell [11]. The Raw percent (K) is widely regarded as too unpredictable to be used for crossover signals and expressed in Equation 2.10.

$$\% \text{ K} = 100 \times \frac{\text{Close -LowestLow (last n periods)}}{\text{HighestHigh (last n periods)- LowestLow (last n periods)}} \tag{2.10}$$
$$\%\text{D}$$

Stochastic Momentum Index (SMI):examines a modification of the stochastic oscillator. It is a more reliable indicator that produces fewer false swings. It calculates the difference between the current closing price and the high/low range of the price range. The SMI typically has a value range of +100 to -100. The SMI, like the stochastic oscillator, is generally employed by traders or analysts to signal overbought or oversold levels in a market [11]. SMI is calculated using Equation 2.11.

$$\text{SMI} = 100 \times \frac{\text{cm}}{\frac{\text{h}}{2}}, \text{ Signal } = \text{EMA}(\text{SMI}) \tag{2.11}$$

WilliamR– Williams Percentage Range (WPR): is quite similar to RSI and Stochastic, and hence, it should not be used in combination with them. It is utilized to determine oversold and overbought levels, as well as the best entry points. Williams % R is set to 14 periods by default, which can be days, weeks, months, or an intraday timeframe. The values of the indicators range from -100 to 0. Values less than -80 indicate that the item has been oversold and will most likely increase in value. Values greater than -20 indicate that the asset has been overpriced and will most likely lose value [11]. Equation 2.12 shows the mathematical formulation of the WilliamR.

$$\%\text{R} = 100 \times \frac{\text{HighestHigh}}{(\text{ last n periods }) - \text{ clos e}} \text{HighestHigh}_{\text{ (last)}} - \text{LowestLow}_{\text{ (last n periods)}} \tag{2.12}$$

Chande's Momentum Oscillator (CMO): is a variant of the RSI. The CMO divides overall movement by net movement ((up-down) / (up + down)), whereas the RSI divides upward movement by net movement (up / (up + down)). Values more than 50 suggest a purchase signal, while values less than 50 indicate a sell signal [11]. Chande's Momentum Oscillator can be given by Equation 2.13.

$$\text{CMO} = 100 \times \frac{\text{ups}_{\text{ i}}\text{-downs}_{\text{ i}}}{\text{ups}_{\text{ i}} + \text{downs}_{\text{ i}}} \text{ where } \quad ups_{\text{i}} = \sum_{\text{i}-\text{n}}^{\text{i}} \text{up and downs}_{\text{ i}} = \sum_{\text{i}-\text{n}}^{\text{i}} \text{down} \tag{2.13}$$

Commodity Channel Index (CCI): shows cyclic trading patterns and detects trends start and end. The range is between 100 and -100; numbers outside this range indicate overbought or over-sold events. If the price increases but the CCI does not, a reversal of the tendency is possible [11]. CCI is calculated using Equation 2.14.

$$\text{CCI} = \frac{\text{TP} - \text{ATP}}{0.015 \times \text{MD}}, \text{ Typical Price (TP)} = \frac{\text{close } + \text{ high}_{\text{ n}} + \text{ low}_{\text{ n}}}{3} \tag{2.14}$$

2.2.3 Volatility Indicator

Volatility indicators provide useful information about the range of buying and selling that take place in given market. The indicator helps the traders determine potential points where the market may change direction [11]. The following are volatility indicators:

Bollinger Bands (BBands): are divided into three lines. The center band represents the simple moving average of the typical price (TP). Upper and lower bands are defined by F standard deviations (typically 2) above and below the center band. The price volatility is wider and narrow accordingly if the price volatility is higher or lower. Bollinger Bands do not provide buy or sell signals but rather serve as a warning indicator for overbought or oversold conditions. When the price approaches the bands above or below, it indicates that the trend could change. The medium band acts as a resistance measure [11]. The upper and lower bands can also determine price objectives as shown in Equation 2.15.

$$\begin{aligned} &\text{MidBand} = \text{SimpleMovingAverage(TP)} \\ &\text{UpperBand} = \text{MidBand} + \text{Price} \times \sigma(\text{TP}) \\ &\text{LowerBand} = \text{MidBand} + \text{Price} \times \sigma(\text{TP}) \end{aligned} \tag{2.15}$$

Average True Range (ATR): is the true range's moving average. The ATR is a measure of volatility. Higher numbers imply high volatility, whereas low values suggest little volatility and a flat price [11]. Equation 2.16 shows the mathematical formulation of the ATR.

$$ATR = \frac{TR_{(-1)} \times (n-1) + TR}{n} \text{ where } TR = \text{TrueHigh} - \text{TrueLow} \tag{2.16}$$

2.2.4 Volume Indicator

Volume is the amount of trading activity that occurs within a determined interval, regardless of price. When volume levels rise beyond their average, it signals the deepening of a trend or confirmation of a trade direction [11]. The biggest trends frequently emerge as volume grows, resulting in substantial price movements.

Chaikin Money Flow (CMF): compares total volume over the last n periods to total volume multiplied by the Closing Location Value (CLV) during the same period. The CLV is calculated by determining where the issue closes within its trading range. When the money flow exceeds 0.25, it is a bullish signal; it is a bearish signal when it falls less than -0.25 [11]. CMF can be used to calculate buying and selling pressure can be given by Equation 2.17.

$$\text{CLV} = \frac{(\text{ close} - \text{low}) - (\text{high} - \text{close})}{(\text{high} - \text{low})} \text{CMF}_{\text{i}} = \frac{\sum_{\text{i}-\text{n}}^{\text{i}} (\text{CLV} \times \text{volume})}{\sum_{\text{i}-\text{n}}^{\text{i}} \text{volume}} \tag{2.17}$$

On Balance Volume (OBV): is the sum of the up and down volume. When the close exceeds the previous close, the volume is added to the running total; when the close falls below the previous close, the volume is reduced from the running total. If the price moves before the OBV, then it is a non-confirmed move. In the OBV, a series of rising or declining peaks indicate a strong trend [11]. If the OBV is flat, the market is not trending. Equation 2.18 shows the mathematical formulation of the OBV.

$$\begin{aligned} IFClose > Close_{[-1]} then OBV = OBV_{[-1]} + Volume \\ ElseIfClose < Close_{[-1]} then OBV = OBV_{[-1]} - \mathrm{V}olume \\ ElseOBV = OBV_{[-1]} \end{aligned} \tag{2.18}$$

Money Flow Index (MFI): calculates the ratio of money flowing into and out of a security and its values range from 0 to 100. Values above 80/below 20 indicate market tops/bottoms [11]. MFI is calculated using Equation 2.19.

$$\begin{aligned} \text{MoneyRatio}_{\,i} &= \frac{\sum_{i-n}^{i} \text{PositiveMF}}{\sum_{i-n}^{i} \text{NegativeMF}} \text{ and} \\ \text{MFI} &= 100 - \left(\frac{100}{1 + \text{MoneyRatio}} \right) \end{aligned} \tag{2.19}$$

2.3 Different Machine Learning Algorithms

2.3.1 Extreme Learning Machine (ELM)

ELM was primarily used to boost the efficiency and speed of single-hidden-layer feedforward networks (SLFNs). The ELM method does not require hidden nodes/neurons to be adjusted, contrary to popular perception in neural network generalization theory, linear theory, and control theory. Unlike ANN, which assigns hidden nodes regularly, ELM allocates hidden nodes at random, creates biases and input weights of hidden layers, and determines output weights using least-squares techniques. The model structure of ELM is shown in Figure 2.1, with j input layer nodes, n hidden layer nodes, m output layer nodes, and the hidden layer activation function g(x). the outputs of the hidden layer can be expressed as (Equation 2.20), and the numerical relationship between output of the hidden layer and output of the output layer can be expressed as (Equation 2.21):

$$h = g(ax + b) \tag{2.20}$$

$$h\left(x_i\right) V = y_i, \quad i = 1, 2, \ldots, N \tag{2.21}$$

The above equation can be written compactly as

$$HV = Y \tag{2.22}$$

Where

$$H = \begin{bmatrix} g(\vec{a}_1, b_1, \vec{x}_1) & g(\vec{a}_1, b_1, \vec{x}_2) & \cdots & g(\vec{a}_n, b_n, \vec{x}_N) \\ g(\vec{a}_2, b_2, \vec{x}_1) & g(\vec{a}_2, b_2, \vec{x}_2) & \cdots & g(\vec{a}_n, b_n, \vec{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ g(\vec{a}_n, b_n, \vec{x}_1) & g(\vec{a}_n, b_n, \vec{x}_2) & \cdots & g(\vec{a}_n, b_n, \vec{x}_N) \end{bmatrix}^T \tag{2.23}$$

$$V = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix}_{n \times m}, \quad Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix}_{N \times m}, \tag{2.24}$$

where $a_i = [a_{i1}, a_{i2}, \ldots, a_{in}]^T$ are the weights connecting the ith input nodes and hidden layer, b_j is the bias of the jth hidden node, and $v_j = [v_{j1}, v_{j2}, \ldots, v_{jn}]^T$ are the weights connecting the jth hidden node and the output layer. V is output matrix of the neural network. We need to set input weights a_{ij} and the bias of the hidden layer b_j; the output weights V can be obtained by a series of linear equations transformations. In conclusion, using ELM to obtain the output weights V can be divided into three steps. Step 1. Randomly select numerical values between 0 and 1 to set input weights a_{ij} and the bias of the hidden layer b_j. Step 2. Calculate the output matrix H. Step 3. Calculate the output weights V:

$$V = H^{\dagger} Y \tag{2.25}$$

where $H^{\dagger}$ represents the generalized inverse matrix of the output matrix H.

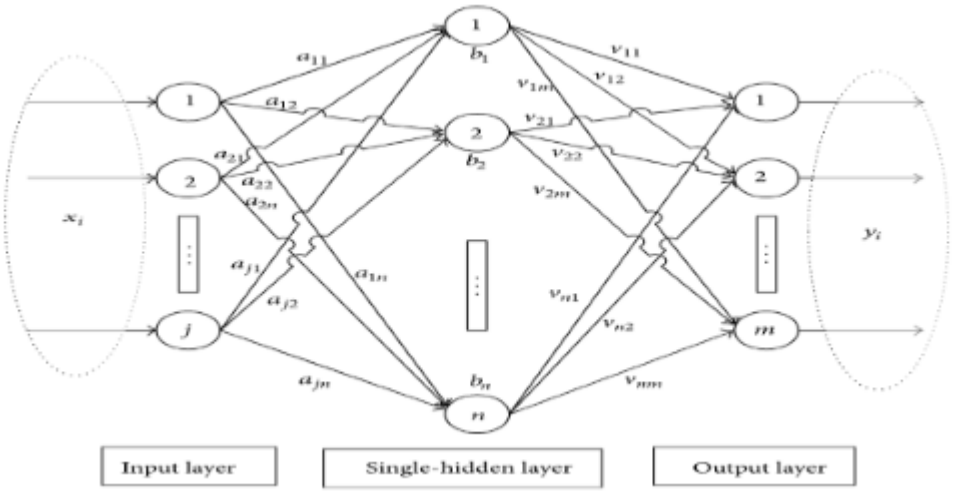

FIGURE 2.1: The working principle of Extreme Learning Machine [19].

2.3.2 Random Forest Algorithm

Random Forest is an algorithm that uses the concept of ensemble methods to solve problems of classification, regression, and feature selection. This algorithm deploys the theory of bagging to train several decision trees in parallel for carrying out the previously mentioned basic tasks. Bagging involves two main processes: bootstrapping and aggregation. Bootstrapping and aggregation are done sequentially. During the bootstrapping step, numerous decision trees are trained using distinct subsets of the training dataset that consist of different subsets of characteristics accessible in the training dataset. This makes the individual decision trees in a random forest unique. Then the aggregation process will follow aggregates of the individual decision trees for the final decision. The random forest algorithm's final decision will be determined by majority voting. Majority voting involves taking the result of the selection of a majority of individual decision trees (in the case of classification problems) as a final result or through averaging, which takes the mean value of the individual decision trees output as a final decision (regression problems) [20]. The basic working principles of the RF algorithm are shown in Figure 2.2.

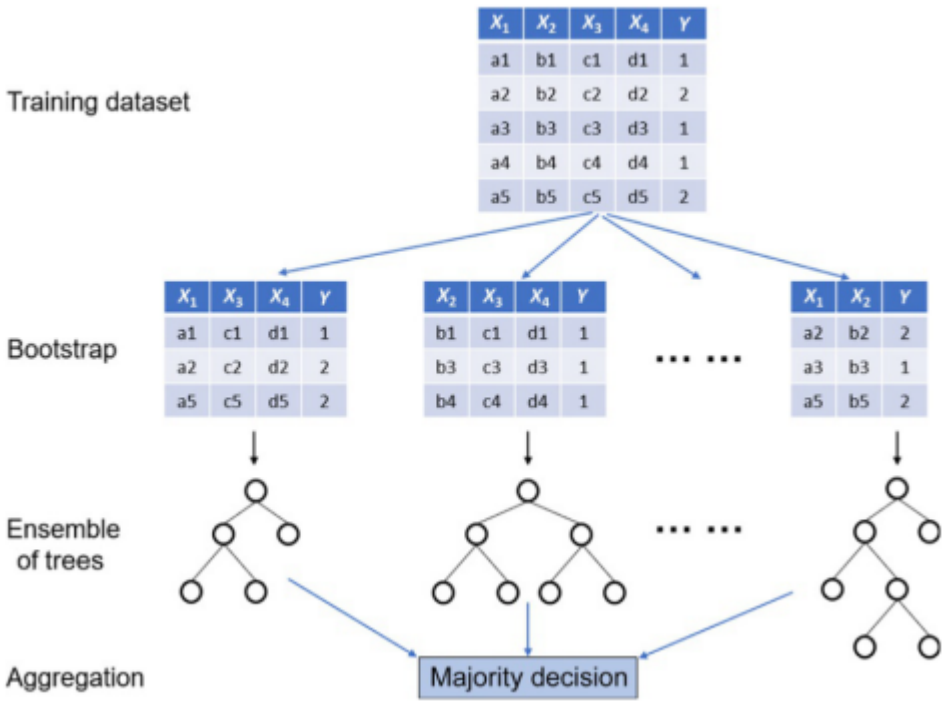

FIGURE 2.2: The working principle of Random Forest algorithm [20].

2.3.3 Artificial Neural Networks

An artificial neural network (ANN) is an ideally linked set of nodes to represent the network of neurons in the brain [21]. ANN is widely used in literature because of its ability to learn complex patterns. The artificial neural network is comprised of nodes (shown as circles in Figure 2.3), an input layer represented as input1 . . . , input4, an optional hidden layer, and an output layer y. ANN aims to determine a set of weights w (between the input, hidden, and output nodes) that minimizes the total sum of squared errors. During training, the weights are modified based on a learning parameter [0, 1] until the outputs match the output. Large values of lambda may result in overly abrupt changes to the weights, while small values may need more repetitions (called epochs) before the model learns adequately from the training data.

The difficulty of using artificial neural networks is finding parameters that learn from training data without overfitting (i.e. memorizing the training data). If there are too many hidden nodes, the system may over fit the current data, while if there are too few; it can prevent the system from properly fitting the input values. In addition, a stopping criterion has to be chosen. The stopping criteria could be based on the total error of the network falling below some predetermined error level [21].

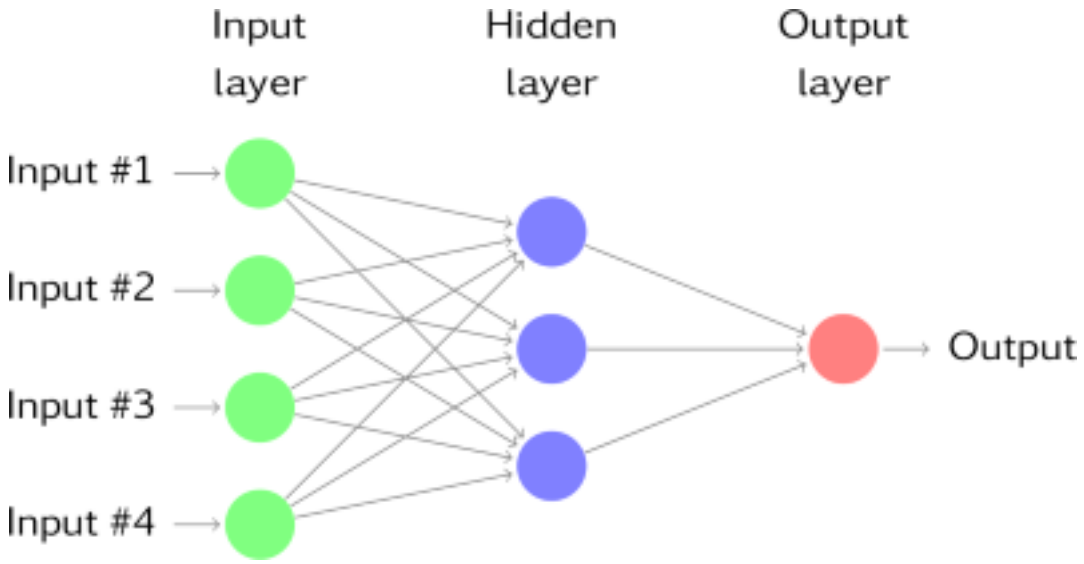

FIGURE 2.3: The working principle of Artificial Neural Network [21].

2.3.4 Support Vector machine

Support vector machines are supervised learning models that can be used for classification, prediction, and clustering problems. A Support Vector Machine takes a set of input observations and constructs a model that can classify new observations into one class

or the other. The model consists of a mapping of the training observations as points in space; it separates the observation sets by dividing them into two classes. This linear separation has a wide area surrounding it. A linear partition in a higher-dimensional space by a hyper-plane corresponds to a nonlinear partition in the output space. This higher dimensional partitioning is known as the "SVM kernel" and can be defined by any mathematical surface. Some example kernels are linear, quadratic, polynomial, and Gaussian radial basis functions [14].

The SVM tries to separate the data into distinct groups using a hyperplane. In the case of two dimensions, a simple line can be drawn to separate the data. A line that is bold separates the data. The distance from the centerline to the sidelines shows the margins.

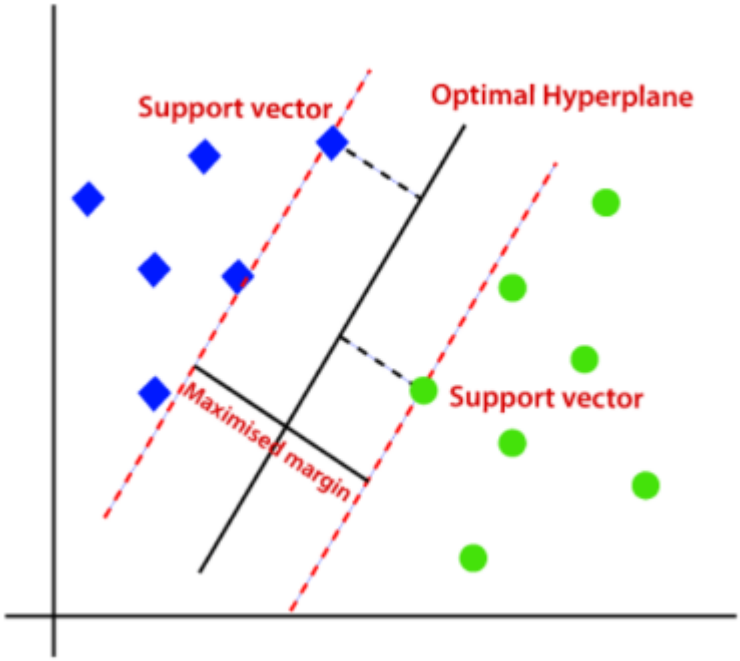

FIGURE 2.4: The working principle of Support Vector Machine [14].

An SVM places the bold line as far away as possible from the nearest observations of the two classes, thus maximizing its margin. Each observation becomes a constraint that needs to be considered when looking for the line solution. The instance located farthest away from the centerline is the correct forecast. In the case of real-world applications, it is not usually possible to get a line that perfectly separates the data within the space. As a result, we may need to employ a curved decision boundary. It is feasible to obtain a hyper-plane that can split the data, although this may not be desired if the data contains noise. In such instances, the soft margin approach [14] must be used [14]. The sloppy margin method allows the points to appear on the wrong side of the margin. These points

have a negative impact. The penalty increases as the distance between the penalty points and the margin increases.

2.4 Feature selection Algorithm

Feature selection is critical in the data analysis process since it extracts useful and non-redundant characteristics. It is the selection of a subset of the original input variables. The chosen features had better capture the properties of the original dataset, and prediction with these features can enhance accuracy [22], [23]. Choosing essential market elements has more influence on forecasting market pricing. Hence, we aim to employ feature selection algorithms.

2.4.1 Correlation-based Feature Selection (CFS)

Correlation-based feature selection [22] is used to handle feature redundancy. The feature selector is easy and quick to execute. The method removes unnecessary and redundant data and, in many situations, enhances the performance of machine learning algorithms. The technique also produces results comparable with a state-of-the-art feature selector from the literature but requires much less computation. The approach examines the correlation between the attribute and the class with the notion that an ideal collection of characteristics should be highly correlated within the class while being uncorrelated with other features. This is done to reduce redundancy and the number of features. This hypothesis is based on two components: the usage of individual variables for prediction and the numbers of intercorrelation among them. It is possible to express it as follows:

$$\text{merits} = \frac{\mathrm{Kr_{cf}}}{\sqrt{\mathrm{K} + \mathrm{K}(\mathrm{K}-1)\mathrm{r_{ff}}}} \tag{2.26}$$

Where Merits is the ”merit” of a feature subset S containing k features:

$$r_{ff} = \sum_{fi \in S} \frac{1}{K} \sum (fi, c) \tag{2.27}$$

Is the mean feature-class correlation (f E S, and c is the class), an indication to how easily a class could be predicted based on the feature; and rff is the average feature intercorrelation between the features which indicates the level of redundancy between them. Feature correlations are measured via information gain that determines the degree of association between features. The information gain (IG) of feature X to the class Y can be expressed in the following Equations:

$$IG(X, Y) = H(X \mid Y) \tag{2.28}$$

$$H(X) = -\sum_{i} P\left(x_i\right) \log_2 \left(P\left(x_i\right)\right) \tag{2.29}$$

$$H(X \mid Y) = -\sum_{j} P\left(y_j\right) \sum_{i} P\left(x_i \mid y_j\right) \log_2 \left(P\left(x_i \mid y_j\right)\right) \tag{2.30}$$

CFS uses the best first (BF) search to explore the search space. It evaluates the merit of a feature by estimating its predictive ability and the redundancy it introduces to the selected feature set. Specifically, this algorithm calculates feature-class and feature-feature correlations first and then selects a subset of features using the Best First search with a certain stopping criterion. It chose the most significant attributes while avoiding reinventing duplication to the maximum extent possible. The CFS model does not need to reserve any training data for the subsequent evaluation. Besides, it works well on smaller data sets.

Chapter 3

Literature Review

Time-series prediction has been used in many practical applications, such as financial forecasting and agricultural price forecasting. According to previous research, time series prediction approaches can be categorized into three types: (1) traditional forecasting method, (2) intelligence forecasting methods, and (3) hybrid models that integrate the first two methods [24]. Figure 3.1 illustrates how the literature is organized in this chapter.

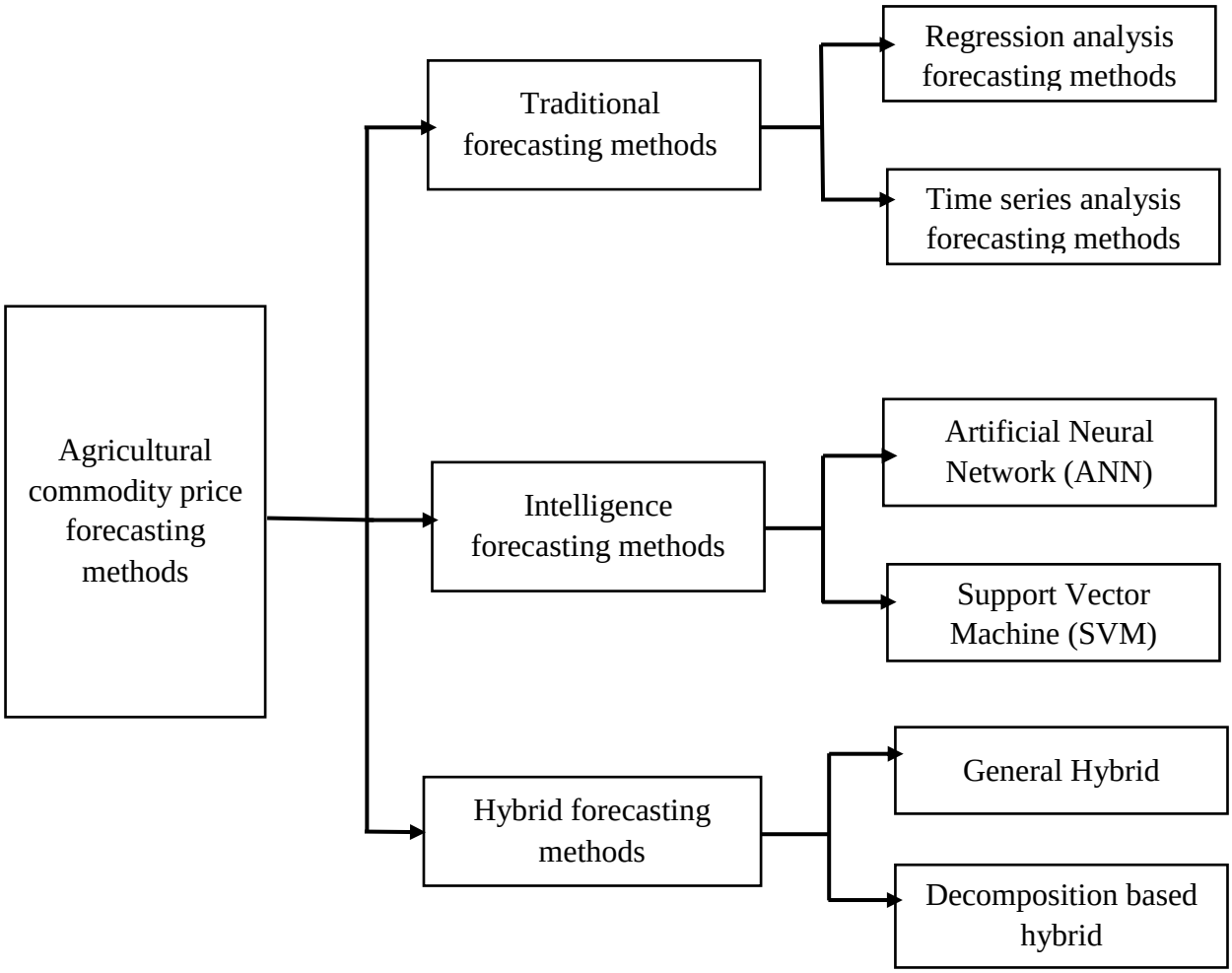

FIGURE 3.1: Literature Review Framework [24].

3.1 Agricultural Commodity Price Prediction Using Traditional Method

Statistical forecasting method is the traditional method of agricultural commodity futures price forecasting and deals with the basic properties of commodity but ignores other critical features such as stationarity and nonlinear behavior of agricultural commodity. These forecasting methods are categorized into two groups namely: time series analysis forecasting methods and regression analysis forecasting methods.

3.1.1 Regression Analysis Forecasting Method

Regression analysis is working based on an analysis of the price-influencing-factor relationship to the price-influence-factor equation. Moore initially proposed correlation- and linear regression patterns to predict cotton prices, and it shows better accuracy than traditional qualitative studies in the regression model. A model such as exponential smoothing (ES), single exponential smoothing, quadratic exponential smoothing, and Holter-Winters seasonal exponential smoothing methods are typical.

Ge and Wu [4] attempted to create a regression analysis to forecast the price of maize. The results showed that the prediction accuracy was low and could only be used for rough forecasts. Although considered the main effect of the supply-demand relationship on maize price changes, the selection of the factors influencing price changes was still incomplete, and the accuracy was somewhat general.

Hyndman et al. [25] formalized the 30 exponential smoothing models by abbreviating the three components: Error, Trend, and Seasonality (ETS). As a result, Holt's exponential smoothing model is denoted by AAN, which stands for additive error, additive trend, and no seasonality. In contrast, the MAdM model has a multiplicative error, damped additive trend, and multiplicative seasonality. Evans and Nalampang [26] employed a multivariate regression model to forecast the price trend of U.S.avocado.

Agricultural commodities data usually has non-linear and non-stationary properties that make the forecasting of commodities challenging. Several price prediction studies that employ statistical approaches treat the data as if it were linear [2]. This assumption contradicts the actual nature of the market, and hence, affects the result.

3.1.2 Time Series Analysis Forecasting Method

Time series analysis is one of the widely used methods to forecast commodity prices. This analysis ignores the influence of every single non-time element on the time series and the

development of various factors as an unavoidable phenomenon on the time axis. This approach has a significant benefit in the case of complex systems. However, the model's interpretation ability suffers because it ignores price influence features. This approach is appropriate for continuously forecasting things and this strategy is significant for forecasting results in a continuous process. It must arrange data from several years into a chronological sequence, and its trends and correlations must be clear, stable, and strong.

A model such as autoregressive moving average (ARIMA), generalized ARIMA, and SARIMA are typical example of time series algorithms [3], [5]. Among the three models, SARIMA model is the most prevalent method for forecasting seasonal agricultural commodities [3]. However, the prediction performance may be lower when the SARIMA is created with linear assumptions for non-linear time series data.

Bv and Dakshayini [17] proposed the Holt Winter model to forecast vegetable price and demand. The performance of Holt Winter's model is compared with the baseline models' linear and multiple linear regression models. The finding indicated Holt Winter's seasonality model has the best performance. Darekar and Reddy [27], Jadhav et al. [5], and Pardhi et al. [28] forecast commodity using ARIMA methods.

Assis and Remali [29] figure out the forecasting performance of four univariate time series methods for Bagan Datoh cocoa bean short-term price forecasting. They used mean absolute error and relative mean squared error to assess the performance of the time series method. The experimental results show that the generalized ARIMA model performed well when compared to other time series methods.

3.2 Agricultural Commodity Price Prediction Using Intelligence Method

The traditional (statistical) forecast method works well to address general linear problems. However, sometimes the value of the market price of agricultural commodities contains non-linear patterns so that predictions using the traditional methods are unstable. Intelligent predictions are a way of predicting future price movements using artificial intelligence. Artificial neural networks and support vector machines are the most common used approaches in agricultural commodity price forecasting.

3.2.1 Artificial Neural Network (ANN)

Artificial Neural Network consists of several simple interconnected neurons. It reflects many fundamental features of human brain activity and is a very sophisticated nonlinear system of dynamic learning. Because ANN has the strong nonlinear computational ability and the ability to approximate a nonlinear classifier with good precision, it can be used for price prediction. Consequently, it is capable of capturing complex nonlinear modes. ANN can also perform massive parallelism, distributed processing, selforganization, self-adaptation, and self-learning.

Li et al. [30] developed a convolutional neural network by combining chaos theory and neural network to predict agricultural commodities such as Pork, Egg, and Potato. The proposed model has a better nonlinear fitting ability and higher precision over linear models for forecasting the agricultural product price. Absolute error (AE) and root mean error (MAE) are used to evaluate the performance of their proposed method.

Recently, Huang et al. [19] proposed another type of ANN, namely an extreme teacher (ELM). The ELM network is a single-layer, more efficient single layer feed neural(SLFN) network with improved study algorithms. The most intriguing benefit of ELM is its relatively fast learning speed, which is faster than traditional ANN and comparable to SVR. The generalization of ELM is very high. Furthermore, unlike traditional ANN, only the ELM output weights are defined by random determination of the input weights and bias [19]. In several areas including, agricultural commodity price forecasting [12], energy market forecasting [31], and stock market analysis [32], the ELM has received considerable attention from the academic community.

A neural network model can easily track and predict complex nonlinear modes. The depth of study and the broad implementation of these models gradually show shortcomings, such as a slow convergence speed, and easily falling into the local minimum. Scholars have used genetic algorithms (GA), Particle Swarm Optimization (PSO), and other optimization methods to improve the model's predictive accuracy in dealing with these challenges. Combining cross selection and mutation selection, the GA improves the neural network weights and thresholds. It efficiently eliminates the drawbacks of neural network overfitting by moving from the local-optimal solution to the global-optimal solution [7]. The PSO optimizes the neural network's weights and thresholds, which increases the model's convergence speed [33]–[35].

3.2.2 Support Vector Machine (SVM)

SVM is a machine learning method that focuses on small sample data learning rules. It is categorized into two: Support Vector Regression and Support Vector Classification, with SVR being widely used for numerical forecasting Wang et al. [36]. SVM generates the ideal decision hyperplane in high-dimensional feature space based on the principle of reducing structural risk and then predicts agricultural product prices. It performs well when dealing with nonlinear sequences and has a high degree of generalization. The support vector machine (SVM) is designed to forecast the nonlinear component of the garlic price series [36]. To resolve these concerns, researchers have employed optimization strategies to streamline SVM penalty and kernel function parameters.

Zang et al. [37] Transforming time-series information from the price index of the agricultural commodity into Low, R, Up, and fuzzy information granulation particles for indicating price movement trend and amplitude, establishing the MEA-optimized price index RSV to forecast price fluctuation and trend change. The empirical investigation has shown that MEA-SVM is more accurate in predicting and takes less computational time.

Wang et al. [38] developed SVM model to predict the nonlinear component of the garlic price series. These studies produced good results and demonstrated the rationality of SVM when applied to the field of agricultural product price forecasting. The studies also revealed SVM's shortcomings, the need for a large amount of calculation, and sensitivity to parameters. Scholars have used optimization algorithms to optimize SVM penalty parameters and kernel function parameters to solve these problems. They created SVM to forecast the prices of commodities: Duan et al. [39] used the time series method to test the stationarity of the aquatic product price series and determine the relevant order, then built the GA optimized SVR to forecast future prices. The results revealed that the error was small.

3.3 Hybrid Forecasting Method

Several theories and practices have demonstrated that the linear and non-linear aspects of commodity price series cannot be captured concurrently by a single component. Consequently, researchers [38] developed a hybrid model for predicting agricultural commodity prices. The hybrid model is a method that combines the results of different approaches of prediction in order to obtain new forecast results. The most popular combination idea is to construct the weight with the well-known regression. In past time-series literature, the assumption has been confirmed. Here we categorize the hybrid model into two subgroups: a general hybrid model and a hybrid model based on the decomposition method.

3.3.1 General Hybrid Model

The principle of this hybrid model is based on the premise that the agricultural commodity price series split horizontally into two parts and that the two parts predict independently using various methodologies before predicting jointly.

Instead of using a single model, general hybrid models that integrate methods such as time-series preprocessing and optimization are frequently used in agricultural price forecasting research. Luo et al. [40] Four models were proposed to forecast Lentinus edodes prices. The models were built using four data mining algorithms: RBF Neural Network BPNN, a NN based on genetic algorithm, and an combined model. The efficiency findings revealed that the BPNN performs the worst, the NN based on the GA model performs better than the RBF, and the integrated model achieves the best.

Zhang et al. [37] proposed a quantile regression neural network (QR-RBF) model to predict the price of soybeans in China. In addition, the model performance was improved by using gradient descent with a genetic algorithm to optimize it (GDGA). This finding was also consistent with prior researches [36].

Ran et al. [41] proposed an attention mechanism-based LSTM to predict travel time. The attentionbased LSTM model proposed in various experiments outperformed other baseline models, and the attention mechanism was able to focus well on the differences in input features. Yoo [42] used the vector autoregressive method and the Bayesian structured time series model to forecast the price of Korean cabbage. Climate and production factors, as well as price trends and seasonality, were considered. The importance of meteorological data was also take in to consideration because Korean cabbage is a crop grown in open fields.

3.3.2 Decomposition-Based Hybrid Model

A decomposition-based hybrid model is different from the general hybrid model. First, it vertically breaks down the prices of agricultural products into many time series components and then predicts them separately using the appropriate time series prediction algorithms. The integration will generate an overall understanding of the market value and obtain forecast results. Therefore, a promising 'decomposition and ensemble' concept is developed to strengthen the prediction capacity of the existing models based on general methods for decomposition, including Wavelet Analysis(WA) Reboredo and Rivera-Castro [43]. Empirical Mode Decomposition(EMD) Zhang et al. [44].

Xiong et al. [45] proposed a hybrid seasonal trend decomposition using Loess (STL) with extreme learning machine(ELM) method for forecasting agricultural commodities such as

cabbage, tomato, kidney bean, cucumber, and hot pepper prices in China. In this study, time-series data is preprocessed using the seasonal and trend losses (STL) approach by considering the seasonal characteristics of vegetable and vegetable prices.

Zhang et al. [15] proposed seasonal index adjustment forecasting methods to predict agricultural commodity prices. The methods decomposed price series into seasonal and trend components using seasonal index adjustment (SIA), then predicted trend components using Hybrid Grey Wolf Optimization (HGWO)-SVR, and finally restored seasonal components to predict trend components to obtain the prediction value.

Scholars have used several decomposition methods to decompose the agricultural product price series to predict future prices at the same time. Wang et al. [46] proposed four Hybrid Models based on PSO-optimized BPNN and four decomposing methods: EMD, Wavelet Packet Transform (WPT), and variable Mode Decomposition (VMD) to forecast wheat, corn, and soybean prices. The results showed that all hybrid models combined with the decomposition method were higher than the PSO-BPNN model with VMD being first and the WPT second in increasing the prediction capacity of the PSO-BPNN model. The study applied several decomposition techniques and laterally compared the effect of data processing with more reference value on basis of nonlinear, non-stationary characteristics of the market price of commodities.

Although many studies have explored the area of market movement and price prediction, a direct implementation of these studies' output is not practical for the Ethiopian market. The market characteristics differ from country to country. Research in Ethiopia agricultural commodity price forecasting is vital to reduce the risks of fluctuating prices. The previous study conducted considering in Ethiopia context [13][16] has two main shortcomings. First, when forecasting agricultural commodity prices, periodicity properties like seasonality and trend patterns were not considered. Second, while predicting commodity prices, time-series features of agricultural commodities such as volatility, volume, momentum, trend, and linearity were not considered. Therefore, the objective of this study is to examine the impact of periodicity and other time series characteristics on Ethiopian agricultural commodity price forecasting.

Chapter 4

Proposed Approach

The general architecture of the proposed approach is depicted in Figure 4.1. This architecture consists of six components: data collection, data prepossessing, feature extraction, feature reduction, model building, and prediction. Each of the aforementioned components of the system is discussed in detail from Section 4.1 to 4.6. The proposed approach comprises three primary steps: feature extraction, feature selection, and prediction.

Step1: twenty eight time-series features and technical indicators are extracted, including periodicity, momentum, trend, volatility, linearity, and volume. The best forecast model for the time series will select, by comparing the forecast errors of the four candidate models for each dataset.

Step 2: A feature reduction is performed using the correlation-based feature selection technique to reduce feature redundancy and improve overall classification capability. The correlation-based feature selection algorithm determines the ranking of features based on the information gain, and the algorithm provides the final identified features.

Step 3: Four commonly used machine-learning models, Artificial Neural Network, Support Vector Machine, Extreme Learning Machine, and Random Forest, will use to construct the prediction model in the proposed approach.

The performance of the proposed approach is then evaluated using four evaluation metrics: root mean squared error, mean absolute error, relative absolute error, and root relative squared error. The accuracy of prediction will be calculated using the standard formula. Finally, we evaluate the relationship between the agricultural commodities and the optimal forecast model.

4.1 Data Collection

The Ethiopian Commodity Exchange (ECX) provided all of the data for this study. The information gathered spanned the months of January 2009 to January 2019. Trade date, lowest price (low), highest price (high), volume (ton), opening and closing price of coffee and sesame were all included in the ECX data. The data attributes are converted to various standards of time series data and stock market indices used for prediction to ensure the dataset's relevance.

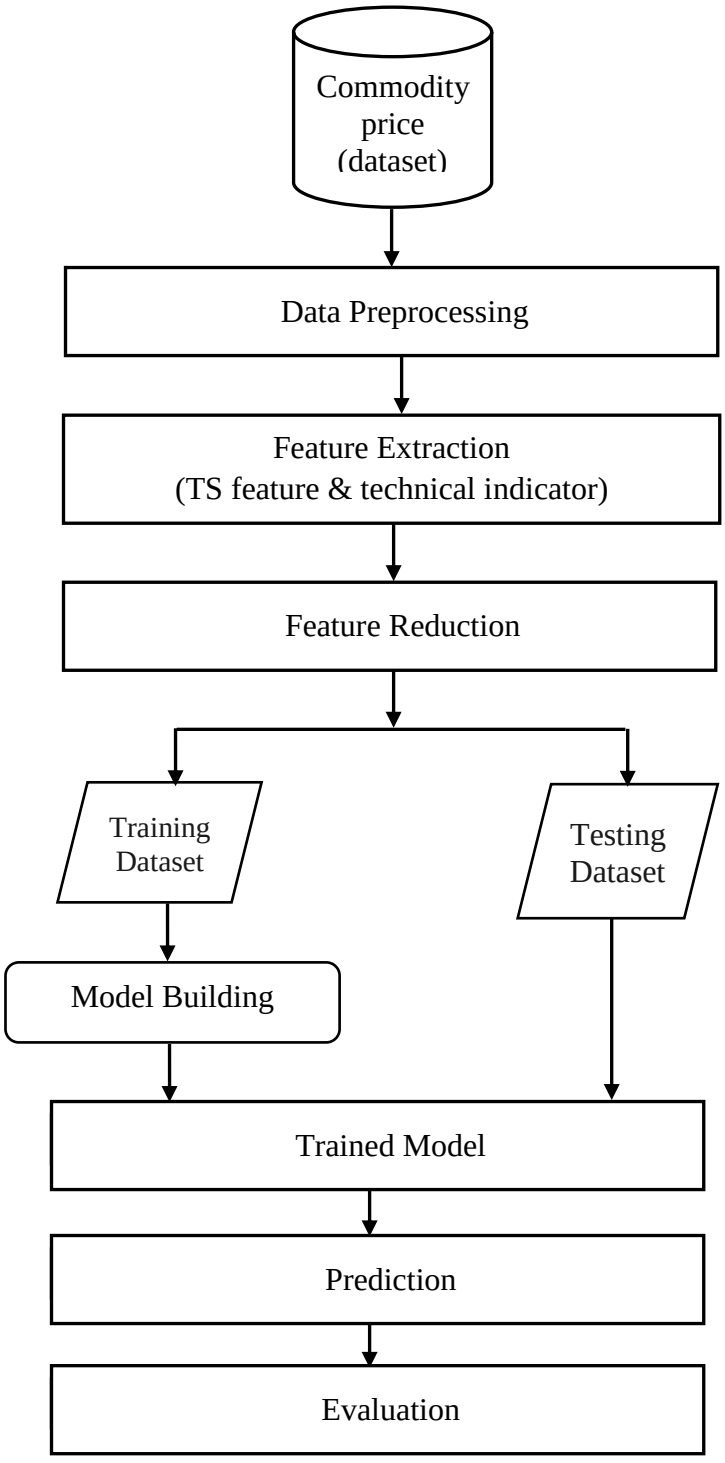

FIGURE 4.1: Proposed Approach.

4.2 Data Preprocessing

Data preprocessing has a considerable impact on the prediction performance of machine-learning models. Several operations are performed in this step to remove noise and create clean data for model training. This stage's efforts have primarily focused on establishing balanced data (to reduce skewing), data transformation and deleting irrelevant data (for instance, those rows that contain null values). The first step in data preprocessing is to determine which features will use in the analysis. Following the selection of the attributes, the data cleaning operation commenced. Finally, data transformation is applied to increase the cleaned data's compliance with the machine-learning model [47].

4.2.1 Handling Outliers

Outliers are extreme values that lie near the limits of the data range (extremely large or small values) or go against the trend of the remaining data. Identifying outliers is important because they may represent errors in data entry. Also, even if an outlier is a valid data point and not an error, certain statistical methods are sensitive to the presence of outliers and may deliver unstable results [47]. There are varieties of methods to handle the outliers that exist in the data; they are eliminating outliers, replace values as missing values, rescaling, and normalization of variables. In this study, the only cause for the existence of outlier is encoding error. To identify these outliers the researcher used data filtering followed by visual inspection methods. The data that are very far large was identified and the records are ignored from the analysis.

4.2.2 Data Cleaning

Data cleaning is a procedure that aims to fill the missing values, smooth out noise while identifying outliers, and fix errors in the data. Witten et al. [47] described data cleaning as a time-consuming and labor-intensive technique required for successful data mining. Some of the data cleaning tasks used in this study include eliminating outliers and dealing with missing values. There are various techniques for avoiding missing values from the final dataset: such as ignoring the tuple value, manually imputing the missing value, or using imputation techniques such as interpolation and mean imputation; replacing the missing value with the same constant value using the mean attribute for all samples of the same class as a specific tuple. The main reason for missing values in the dataset is that the contract may not be traded on a given day. We use the interpolation technique to reduce the impact of the missing value in the dataset). Interpolation is the technique which estimates the value of a function at a point from its values at nearby points.

4.3 Feature Extraction

One of the primary works in prices prediction is to extract valuable features from a variety of financial datasets. Better inputs can contribute to improving the prediction outcomes [11], [12]. In this study, nineteen technical indicators and four time-series features are derived from the dataset prepared in the previous step. Table 4.1 describes the features extracted from the original agricultural commodity dataset (collected dataset). We categorized the features based on the characteristics of agricultural commodity prices into seven major groups: periodicity, volume, volatility, momentum, time-series dataset, and linearity. The details of each feature used in this study are presented below.

4.3.1 Technical Indicator

A technical indicator is a mathematical pattern derived from historical data used by technical traders or investors to predict future price trends. It uses a mathematical formula to derive a series of data points from past price, volume, and open interest data. The following are common technical indicators used in the study.

- **Trend indicators:** are used to measure the direction and strength of a trend using some form of price averaging to establish a baseline. The trend could be positive or negative depending on whether the time-series pattern is up or down.
- **Volume indicators:** represents the amount of trading activity that occurs during a given interval, independent of price. As volume levels move above their average, it indicates the strengthening of a trend or trading direction. The biggest trends frequently occur when volume increases, resulting in significant price fluctuation.
- **Volatility indicators:** provide useful information about the range of buying and selling that take place in a given market information that can help the traders determine potential points where the market may change direction.
- **Momentum indicators:** help to identify the speed of price movement by comparing prices over time. This indicator is usually used for price and volume analysis.

4.3.2 Time Series Features

Time series analysis refers to techniques for evaluating time series data in order to extract relevant data and other data properties such as seasonality and trend.

- **Time series dataset:** contain the original data arranged in time series manner.
- **Linearity features:** are vital to determine the model selection.

- **Periodicity features:** is the fundamental features of time series data and provide indications on periodicity and seasonality of time series.

TABLE 4.1: Features used in this study.

Feature	Description
Time series feature	
Linearity feature	
1. Non linearity	Non linearity of the time series.
2. Linearity	Linearity of the time series.
Periodicity feature	
3. Seasonality	The strength of seasonality
4. Trend	Strength of trend
Technical indicator	
Trend indicator	
5. SMA	Simple moving average gives all the days equal weight.
6. EMA	It gives higher priority to the actual data.
7. DEMA	The price graph closer than mostof other moving Average
8. MACD	It indicates the dynamics and strength of the current trend and oscillates around the zero line in both directions.
9. ARRON	Attempts to show when new trend is dawning.
10. TDI(ADX)	It is used to detect when a trend has begun and end.
Momentum indicator	
11. RSI	It calculates a ratio of the recent upward price movements to the absolute price movement.
12. Stoch	It measures relation between close and recent trade.
13. William%R	help detect oversold and overbought levels, as well as the appropriate entry points.
14. CCI	It shows trading cyclic trends and detect beginning and ending market trends.
15. CMO	The chande momentum oscillator is a modified RSI.
16. ROC	It calculates the percentage difference between two observations in a series.
Volume indicator	
17. CMF	Chaikin money flow indicator
18. OBV	It is a cumulative total of the up and down volume.
19. MFI	It calculates the ratio of money flowing into and out of a security.
Volatility indicator	
20. BBands	Bollinger bands are bands relating to volatility placed below and above a moving average.
21. ATR	It is a measure of volatility.
22. TrueHigh	The true high of the series.
23. TrueLow	The true low of the series
Time Series Dataset	
24. Open	Opening price
25. Close	Closing price
26. High	The highest price given by bidders
27. Low	The lowest price given by bidders
28. Volume	Volume of the commodity provided for bid

4.4 Feature Reduction

Some of the features stated above may capture similar information, resulting in redundancy. These redundancies will increase the complexity of the classifiers and decrease their generalization capability [22], [23]. As a result, feature reduction should be used to remove these redundancies and enhance model selection performance. In this study, correlation-based feature selection is adopted for feature reduction. This approach considers the correlation between features and class besides the redundancy between each pair of two features while selecting a feature set. Correlation based feature selection is utilized in this study to decrease feature redundancy and eliminate the features to enhance overall prediction capabilities. Correlation-based feature selection algorithms determine the feature's ranking based on information gain [22].

4.5 Model Building

After completing the data-preprocessing phase, the next step is to create a model. To build prediction models, we use four commonly used machine-learning algorithms: Artificial Neural Networks, Support Vector Machines, Extreme Learning Machine, and Random Forest. The algorithms are selected based on their strengths and performance in previous studies. As one of main objective of this research work is to figure out the impact of the proposed agricultural commodity price prediction approach, we built separate models for the baseline and proposed approach. We carefully designed the models to have identical settings while building the two independent models, except for some mandatory parameters that must be different.

4.5.1 Artificial Neural Network (ANN)

Artificial Neural Network is a data-driven model that can estimate a wide range of nonlinear problems. [48]. One of the classic neural networks is the back-propagationneural network (BPNN), which includes feed-forward and back-propagation. It is well known for its error-learning algorithm, which adjusts weights and biases. BPNN with a single hidden layer usually gives good accuracy for the time series dataset [33].

4.5.2 Support Vector Machine (SVM)

SVM [49] is based on the risk minimization concept. SVM is used in this study for the following reasons: first, data classification can be performed without making any strong assumptions; second, SVM is based on structural risk minimization theory, which aims to reduce the upper bound of the generalization error and is very resistant to overfitting;

and third, an SVM model is a linearly constrained quadratic program. Its solution is always globally optimum, whereas other models might tend to fall into a locally optimal solution.

4.5.3 Random Forest Algorithm (RF)

Random Forest is an algorithm that solves classification, regression, and feature selection problems using the ensemble approach. This algorithm deploys the concept of bagging to train several decision trees in parallel for carrying out the previously mentioned basic tasks. Bagging involves two primary processes, bootstrapping and aggregation which are done sequentially. During the bootstrapping step, many decision trees are trained using distinct subgroups of the training dataset, which comprises of different subsets of characteristics accessible in the training dataset. This makes the individual decision trees in a random forest unique [20], [50].

4.5.4 Extreme Learning Machine (ELM)

ELM is a single hidden layer feedforward neural networks proposed by [19]. Unlike traditional learning algorithms in feedforward neural network, where parameters are tuned iteratively, the Moore-Penrose generalized inverse is applied to determine the output weights in ELM, thus requiring little time for training. Several advantages of Extreme Learning Machine include the simplicity of usage, quicker speed of learning, greater generalization performance, appropriateness for several nonlinear kernel functions and activation function. This advantage has been applied to classification tasks and regression tasks in numerous studies [12], [31], [32].

4.6 Prediction and Evaluation

After the models are built, then the dataset is divided into two parts for training and testing the respective models. By using the test data, the performance of the trained model is tested using different evaluation parameters, which are extracted, from the standard evaluation metrics. These include mean absolute error, relative mean squared error, relative absolute error, and root relative squared error.

Chapter 5

Experiments

In order to answer our research questions which are defined in Section 1.1, we conducted successive empirical experiments. This section details the characteristics of the dataset used for this research study, the experimental setup and configuration of each experiment, and the evaluation mechanisms applied to measure the performance of the proposed approach.

5.1 Dataset Description

The datasets used for the experiment is collected from Ethiopia Commodity Exchange (ECX). The dataset contains price-related features gathered from January 2009 to January 2019. The features included in the dataset are trade date, lowest price (low), highest price (high), volume (ton), opening and, closing price of coffee, and sesame. The data attributes are converted to time series data and stock market indices used for prediction. Table 5.1 contains an summary of the dataset. Each of the feature used in this study discussed in detail in Section 4.3.

TABLE 5.1: Summary of dataset.

Commodity type	Collection Period	Original dataset	Final dataset for this study	Commodity class	Instance
Coffee	1/2009- 1/2019	33520	29816	10	28
Sesame	1/2009- 1/2019	23212	18560	8	28

5.1.1 Feature Description and Analysis

Some of the features stated in Table 4.1 may capture similar information, resulting in redundancy. These redundancies will increase the complexity of the model and decrease their generalization capability. In this study, the correlation based feature selection is adopted for the feature reduction. The feature correlation diagram based on person matrix is shown in Figure A.1. The light point represents the maximum mutual information value of all the twenty eight features. The greater the correlation, the lighter the color. This study considered features with a correlation value equal to or greater than 0.75 as

strongly correlated. Figure A.1 shows the correlation among features before feature reduction. It can be seen that most of the correlations are dark colored, which reveals that these features contain diverse information on the time series dataset. A few points are light-colored, which implies that the information contained in these features is redundant. These redundancies may have negative effects on the generalization performance of the prediction model. Thus feature reduction should be employed to eliminate the redundancies.

Figure A.2 shows the correlation of the features after feature reduction. Compared to Figure A.1 , the numbers of the light-colored points have been reduced. This result shows that the CFS approach is a workable approach to feature reduction.

After feature reduction, twenty-five features including, five TS dataset feature, four time series feature and sixteen technical indicators, remained. In general, the details of the selected feature for price prediction are listed in Table 5.2.

TABLE 5.2: The final features after feature reduction.

Category	Features
Trend	EMA, DEMA, SMA, MACD, ADX
Momentum	CCI, CMO, ROC, RSI, Stoch
Volatility	ATR, TrueHigh, TrueLow
Volume	OBV,CMF, MFI
Periodicity	Seasonality, trend
Linearity	Linearity, nonlinearity
TS dataset	Close, Open, High, Low, volume

5.2 Experiment Setup

To evaluate the impact of the proposed price prediction, we conducted two sets of experiments. In the first set of experiments, we used the price prediction model computed by [13] to build and evaluate the price prediction models. We refer to the results of these experiments as the baseline. To assess the performance of the proposed approach, we conducted two experiments.

- **Experiment I:** The first experiment aims to assess the impact of seasonality and trend pattern and other properties of agricultural commodity price series such as volatility, volume, momentum, and linearity on the future price prediction. In this experiment, twenty-five features are used, comprising five TS dataset features, four time-series features, and sixteen technical indicators. To test the proposed price

prediction approach, we built a model using four machine learning algorithms: Artificial Neural Network, Extreme Learning Machine, Support Vector Machine, and Random Forest. This experiment aims to demonstrate the proposed approach's ability to predict commodity prices.

- **Experiment II:** The second experiment investigates the individual capability of features based on their contribution level (level of importance) to the prediction model. The total features utilized in experiment 1 for sesame and coffee commodities are computed using a random forest algorithm in this experiment. This experiment aims to show the contribution level of the proposed features in the price prediction scheme.

To assess the performance of our proposed approach, we created a separate training and testing set, with 80% of the data allocated to training and 20% dedicated to testing. These assessment processes are the most prevalent and often used in a wide range of research investigations [47]. To conduct empirical assessments of our proposed approach, we use four machine-learning algorithms: Artificial Neural Network, Support Vector Machine,Extreme Learning Machine, and Random Forest.

5.2.1 Extreme Learning Machine Model Setup

For this algorithm, no special parameter configuration is made. We just used the default model setup. The default hyper-parameter settings are presented as follows.

- Number of Layers: (input layer, single hidden layer, output layer)
- Activation function: rectified linear unit (ReLU) for all layers except the output layer, which is sigmoid.
- Optimizer: Adaptive moment estimation (ADAM)
- Kernel initializer: uniform for all layers
- Loss Function: Mean Squared Error Loss
- learning rate = 0.001
- batchsize = 64

5.2.2 Support Vector Machine Model Setup

We did not focus on hyper parameter optimization specifics for this model because our objective is to assess the amplification impact of the proposed features. We just used the

default model setup. The default hyper-parameter settings are presented as follows.

- kernel = poly-kernel
- regOptimizer = RegSMOImproved
- bachsize = 64
- c = 0
- epsilon parameter = 0.001
- tolerance = 0.001

5.2.3 Artificial Neural Network Model Setup

While selecting the optimum model to conduct experiments, we follow a trial and error scheme as there is no specific formula or guideline to determine the number of layers of the artificial neural network model and the number of neurons in each layer. Therefore, by starting from a simple structure, we try to examine the learning curve of the system and finally, we select the one that gives us a better result.

- Input, hidden, and output layer
- Loss Function: Mean Squared Error Loss
- Kernel initializer: uniform for all layers
- Learning Rate: 0.001
- batchSize: 64

5.2.4 Random Forest Model Setup

For this algorithm, no special parameter configuration is made. We just used the default model setup. The default hyper-parameter settings are presented as follows.

- bachsize = 64
- max-depth = 0 means (unlimited)
- numitration = 100
- Random state = 0

5.2.5 Experiment Environment

The hardware and software specification used to conduct the experiments briefly presented in Table 5.3.

TABLE 5.3: Machine Specification.

Specification of Machine Used For Experiments	
Manufacturer	Dell
Model	Dell Inspiron 5570
Processor	Intel(R) Core(TM) i7-8550U CPU @ 1.8 GHz
Memory	8GB DDR4
Operating System	Windows 10 (64 bit)
Software	R 4.02 and Python 3.9

5.3 Evaluation Metrics

In this study, four popular evaluation metrics were used to see the applicability and performance of the prediction model. Prediction algorithms can depend on different factors such as information quality, class distribution of datasets, and the number of instances. The prediction models performance is assessed using four metrics: the root mean absolute error, mean absolute error, root relative squared error, and relative absolute error.

5.3.1 Mean Absolute Error (MAE)

Mean absolute error calculate the average size of the errors in a series of predictions without consideration of direction It expresses the average model prediction error in units of the variable of interest. Equation 5.1 shows the mathematical formulation of the MAE.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \tag{5.1}$$

Where
Xi is the measurement,
X is the true value.

5.3.2 Root Mean Squared Error (RMSE)

Root mean squared error (RMSE) is the square root of the mean of the square of all of the error. The use of RMSE is very common, and it is considered an excellent general-purpose

error metric for numerical predictions. RMSE measures the difference between sample predicted and observed (actual) price value. RMSE is calculated using equation 5.2.

$$\text{RMSE} = \sqrt{\frac{\sum_{\text{i=1}}^{\text{n}} (\boldsymbol{P}_i - \boldsymbol{O}_i)^2}{\text{n}}} \tag{5.2}$$

Where

Pi is the predicted value,
Oi is the observed value.

5.3.3 Root Relative Squired Error (RRSE)

The root relative squared error (RRSE) represents the difference in what happened when a simple predictor was employed. The fundamental predictor is the average of the actual data. As a result, normalizing the total squared error is as easy as dividing the total squared error by the total squared error of a simple prediction. RRSE is calculated using equation 5.3.

$$E_i = \sqrt{\frac{\sum_{j=1}^{n} \left(P_{(ij)} - T_j\right)^2}{\sum_{j=1}^{n} \left(T_j - \bar{T}\right)^2}} \tag{5.3}$$

Where

P(ij) is the prediction of individual component I for item j out of n samples; Tj is the predicted value for record j.

5.3.4 Relative Absolute Error (RAE)

It is relative to a simple predictor, which is just the average of actual values. It takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor. Equation 5.4 shows the mathematical formulation of the RAE.

$$\delta = \left| \frac{v_A - v_E}{v_E} \right| \cdot 100\% \tag{5.4}$$

Where

$\delta = percenterror$
VA = actual value observed
VE = expected value

5.4 Results

This section presents the details of each experiment result. In order to make the results clear and easy to follow, we present each set of experiments with its associated research question.

5.4.1 Experiment I

Q1. *Does considering time series seasonal and trend pattern of commodities improve price-forecasting performance?*

This research question focuses on assessing the impact of seasonality and trend pattern (periodicity), momentum, volatility, momentum, volume, and linearity features on agricultural commodity price forecasting.

Our experiment results show that the proposed approach is capable of improving the baseline results by an average of 4.3, 4.4, 2.7, and 5.1 while using Artificial Neural Network, Extreme Learning Machine, Support Vector Machine, and Random Forest respectively. The highest improvement in RMSE, 5.1, is observed while using the random forest algorithm. In addition, the proposed approach improves the relative absolute error and root relative squared error.

Among the four machine learning algorithms used in the study, Artificial Neural Network, followed by Extreme Learning Machine and Random Forest, has the lowest score in prediction accuracy while using both proposed and baseline approaches for the majority of the cases. The proposed approach produces consistent MAE values in the Artificial Neural Network, whereas the MAE values in the other models are inconsistent. SVM had the highest MAE improvement of 2.38, followed by RF in sesame, which had an MAE improvement of 1.70. Besides, the commodity sesame has recorded the highest MAE for all models except for Extreme Learning Machine. This shows that the proposed approach have good potential to forecast the future commodity price.

Our proposed approach forecasts the commodity's future price with great accuracy until unanticipated events (covid 19 and political situation) in the commodities market occur. However, the covid 19 pandemic and Ethiopia's political condition make estimating future prices challenging because the commodity selling price has dramatically changed by almost double.

The result demonstrates that the machine-learning algorithm used in commodity price prediction is plays an important role. The comparative results of the baseline and the proposed approach for the four machine learning algorithms are shown from Table 5.4 to Table 5.8.

TABLE 5.4: Result using Artificial Neural Network.

Commodity	Prediction Model	Evaluation Metrics				Delta (Proposed vs Baseline)	
		MAE	RMSE	RRSE	RAE	MAE	RMSE
Coffee	**Proposed**	1.0265	3.0729	0.6837	0.3044	**1.1802**	**4.5098**
	Baseline	2.2067	7.5827	1.687	0.6544		
Sesame	**Proposed**	1.3834	2.2611	0.1984	0.1413	**1.1426**	**4.2408**
	Baseline	2.526	6.5019	0.5705	0.2581		
Average	**Proposed**	1.2049	2.6670	0.4410	0.2228	**1.1614**	**4.3753**
	Baseline	2.3663	7.0423	1.1287	0.4562		

TABLE 5.5: Result using Extreme Learning Machine.

Commodity	Prediction Model	Evaluation Metrics				Delta (Proposed vs Baseline)	
		MAE	RMSE	RRSE	RAE	MAE	RMSE
Coffee	**Proposed**	1.2711	2.8462	0.6332	0.3769	**0.8912**	**4.7768**
	Baseline	2.1623	7.623	1.696	0.6412		
Sesame	**Proposed**	2.2997	4.0513	0.3555	0.235	**-0.787**	**4.0665**
	Baseline	1.5119	8.1178	0.7123	0.1545		
Average	**Proposed**	1.7854	3.4487	0.4943	0.3059	**0.1034**	**4.4216**
	Baseline	1.8371	7.8715	1.2041	0.3978		

TABLE 5.6: Result using Support Vector Machine.

Commodity	Prediction Model	Evaluation Metrics				Delta (Proposed vs Baseline)	
		MAE	RMSE	RRSE	RAE	MAE	RMSE
Coffee	**Proposed**	2.8415	6.8161	1.5165	0.8426	**0.8684**	**1.7926**
	Baseline	3.7099	8.6087	1.9153	1.1001		
Sesame	**Proposed**	6.3596	10.317	0.9053	0.6497	**2.3897**	**3.7068**
	Baseline	8.7493	14.024	1.2466	0.8939		
Average	**Proposed**	4.6005	8.5667	1.2109	0.7461	**1.6290**	**2.7497**
	Baseline	6.2296	11.316	1.5799	0.997		

TABLE 5.7: Result using Random Forest.

Commodity	**Prediction Model**	**Evaluation Metrics**				**Delta (Proposed vs Baseline)**	
		MAE	**RMSE**	**RRSE**	**RAE**	**MAE**	**RMSE**
Coffee	**Proposed**	2.4055	4.6963	1.0448	0.7133	**0.9608**	**4.3878**
	Baseline	3.3663	9.0841	2.021	0.9983		
Sesame	**Proposed**	8.2016	20.094	1.4899	0.8379	**1.706**	**5.9282**
	Baseline	9.9076	26.023	2.2834	1.0122		
Average	**Proposed**	5.3065	10.895	1.2673	0.7756	**1.3334**	**5.158**
	Baseline	6.6369	17.553	2.1522	1.0052		

While doing our experiments to answer this research question, we also try to figure out which algorithm is performing better towards improving agricultural commodity price prediction. As a result, we compare the performance of the four machine algorithms that are used in this research. Here, our basis for comparing one algorithm with another one is based on their RMSE and MAE score in the proposed approach, which relates to different aspects of their predictive capacity.

The experimental results demonstrated that the ANN outperforms the other three algorithms in the majority of the test cases. Based on the findings of the preceding analysis, it is possible to conclude that of the four algorithms used in this research study, the artificial neural network is the best for predicting future prices. The summary of the comparison of the four algorithms are shown in Table 5.8.

TABLE 5.8: Comparative analysis of the applied algorithms.

Prediction model	**Evaluation metrics**	
	MAE	**RMSE**
ANN	1.2049	2.6670
ELM	1.7854	3.4487
SVM	4.6005	8.5667
RF	5.3065	10.8955

To better understand the performance of the proposed approach, we selected two algorithms (Artificial Neural Network and Support Vector Machine) and analyzed ten individual results. The prediction used the proposed price prediction approach as shown a better performance when compared to the baseline approach. The ten days sample price prediction for the two commodities; coffee and sesame, are presented from Figure 5.1 to Figure 5.4.

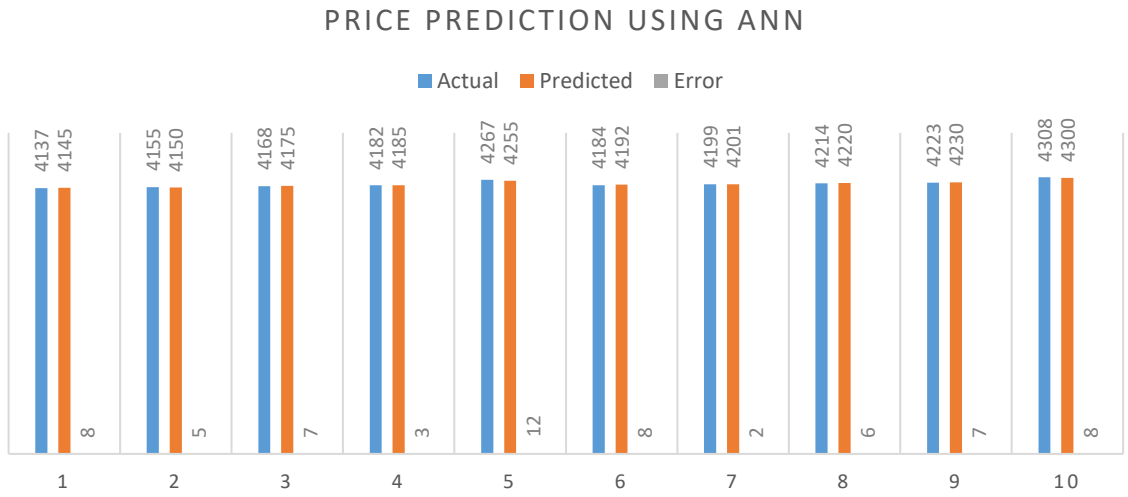

FIGURE 5.1: Sesame price prediction using Artificial Neural Network.

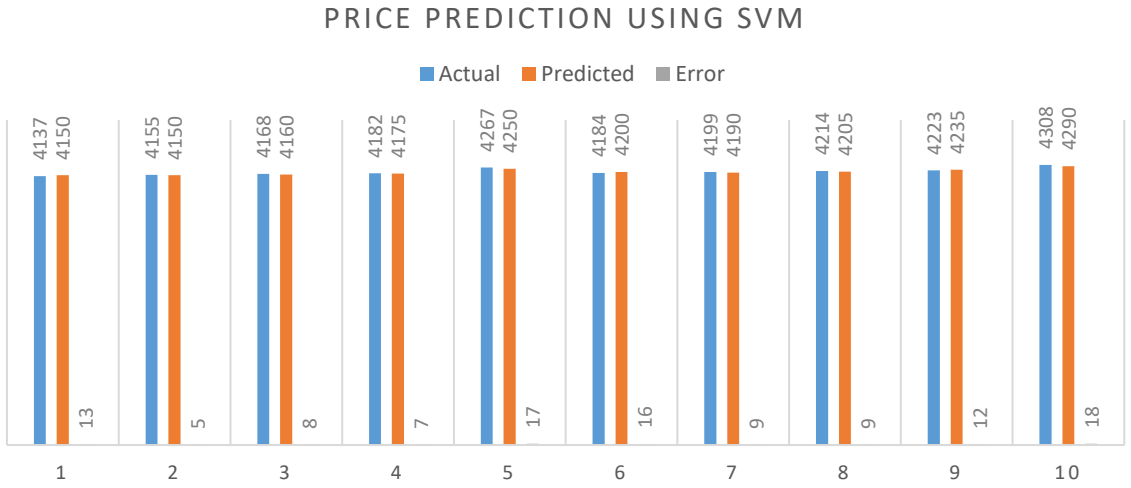

FIGURE 5.2: Sesame price prediction using Support Vector Machine.

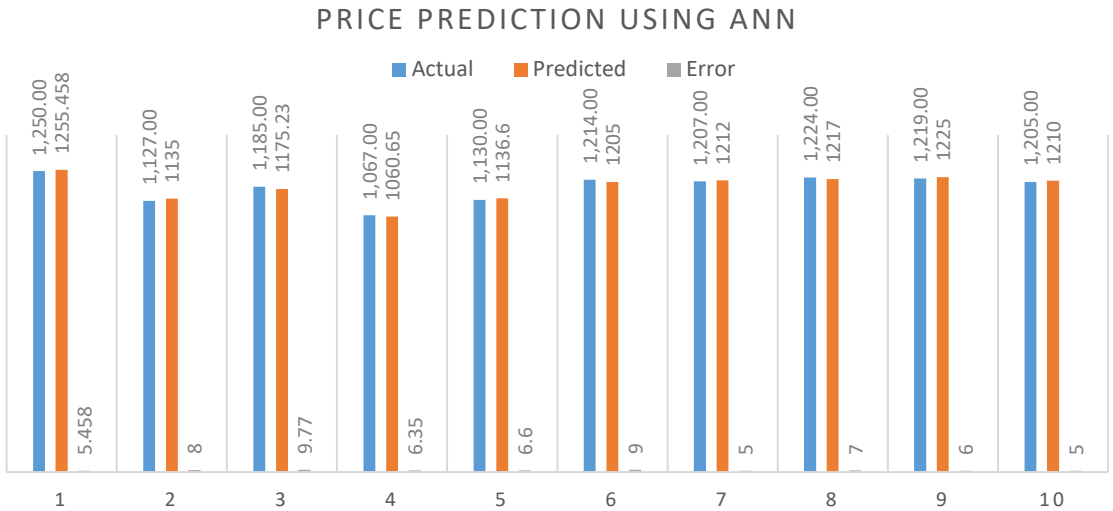

FIGURE 5.3: Coffee price prediction using Artificial Neural Network.

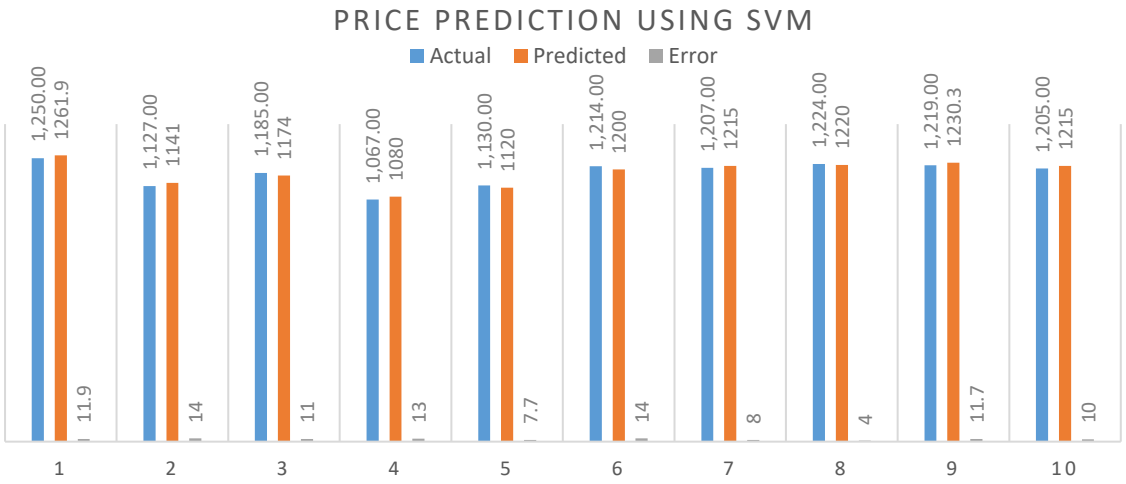

FIGURE 5.4: Coffee price prediction using Support Vector Machine.

Our experiment findings demonstrate that the proposed approach improves the baseline results up to 5.92 in RMSE and 2.38 in MAE. In addition, the proposed approach improves the relative root squared error and relative absolute error. Furthermore, among the four algorithms used in this research work, Artificial Neural Network algorithm outperforms the other algorithms in most of the experiments. The proposed approach forecasts a commodity's future price with higher accuracy in regular market movement. However, the covid 19 pandemic and Ethiopia's political condition make estimating future prices challenging because the commodity selling price has dramatically changed by almost double.

5.4.2 Experiment II

Q2.*With regard to agricultural commodity price prediction, what are the important features in the proposed approach?*

In order to answer this research question, we computed the importance of each feature using random forest algorithms. To evaluate the contribution of each feature for accomplishing the commodity price prediction, we used the random forest algorithms to calculate the relevance of each feature to answer this research question. We compute the importance of each feature for both sesame and coffee price prediction. The random forest algorithm uses a tree-based classifier system to rank the importance of each feature. As shown in the Figure 5.5 and Figure 5.6, we can see that the proposed feature contributed well to the agricultural commodity price prediction scheme.

While random forest algorithms used, for sesame, close, high, low, open, exponential moving average (EMA), double exponential moving average (DEMA), simple moving average (SMA), truehigh, truelow, relative strength index (RSI), average directional movement index (ADX), seasonality, chande's momentum oscillator (CMO), and money flow

index (MFI) are among the top important feature based on their ability to predict future price. In general, the details of feature importance experiment results are presented in Figure 5.5.

In the case of coffee, a similar pattern is observed, with only two features changing in the top important features. The features such as close, high, low, open, double exponential moving average (DEMA), exponential moving average (EMA), simple moving average (SMA), truehigh, truelow, trend, seasonality, Average true Range, relative strength index (RSI), and moving average convergence/divergence (MACD) are among the top important feature based on their ability to predict future price. The overall feature importance chart of coffee commodity are presented in Figure 5.6.

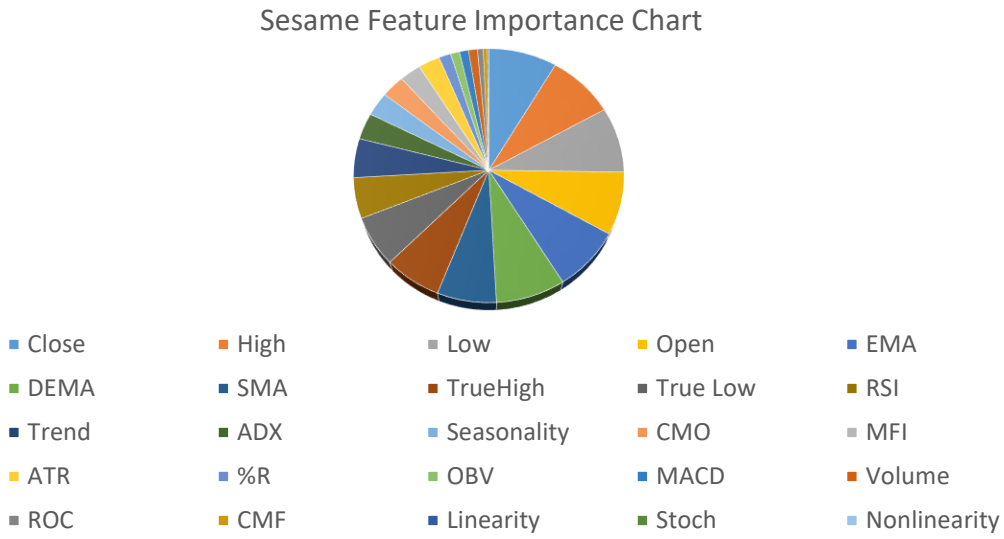

FIGURE 5.5: Sesame feature importance chart using Random Forest.

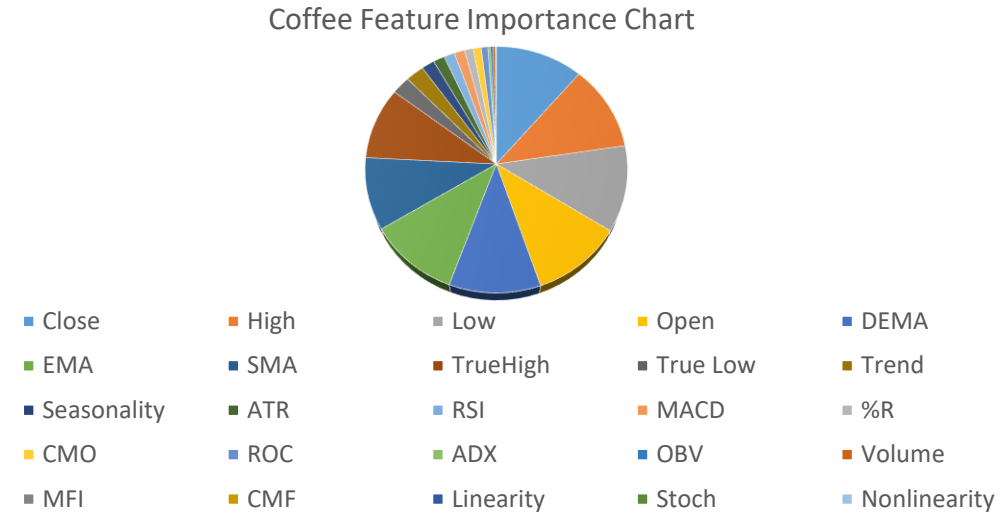

FIGURE 5.6: Coffee feature importance chart using Random Forest.

This experiment ranks each feature based on their contribution to the prediction model. The given features from (Figure 5.5 and Figure 5.6) ranked based on their predictive ability. For the two datasets (sesame and coffee) used in the study, relatively similar results have existed. The high predictive Features such as; close, high, low, open, exponential moving average (EMA), double exponential moving average (DEMA), simple moving average (SMA), truehigh, truelow, trend, seasonality, and relative strength index (RSI) were found in the top important features in their predictive ability. The experimental results confirmed that the features such as periodicity, volatility, momentum, volume, and trend contribute to enhancing price prediction performance.

5.5 Discussion

All sets of experiments conducted using different algorithms and datasets show the potential of proposed approach features such as periodicity, volatility, momentum, volume, trend, and linearity towards enhancing agricultural commodity price prediction. By doing empirical experiments, we come up with three fundamental findings.

The first finding shows that considering periodicity, volatility, trend, momentum, volume and linearity aspects of the commodity into account has the potential to improve the performance of prediction models for the Ethiopian agricultural commodity price prediction. Second, we conducted an experiment to see which of the four machine learning algorithms used in this work perform better. According to the results, random forest improves agricultural commodity price prediction by a significant delta.

Regarding the potential of enhancing the performance of the agricultural commodity price prediction scheme, the proposed features have a good potential. Among eight test cases, seven cases show that the proposed approach improves the performance of the price prediction in terms of root mean squared error and mean absolute error. The proposed approach improved the baseline findings by 5.92 in root mean squared error (RMSE) and 2.38 in mean absolute error (MAE). Furthermore, the proposed approach even was capable to enhance the relative absolute error and the root relative squared error. Based on the findings, we may conclude that considering commodity seasonal and trend characteristics and other time-series characteristics can help the agricultural commodity prices prediction.

One of our objectives of conducting the experiments is to identify a better algorithm, which is helpful to forecast price for further implementation suggestions. In this regard, we compare the results of the four commonly used regression algorithms: Artificial Neural Network, Extreme Learning Machine, Support Vector Machine and Random Forest in terms of their RMSE and MAE. As the results of the experiments show, Artificial Neural Network outperforms the other algorithms in the majority of the experiments. Besides,

Extreme Learning Machine, Support Vector Machine, and Random Forest were ranked in their prediction performance with the respective order.

Our empirical findings also demonstrate that Random Forest is the algorithm that performs the worst in the majority of the experiments. We conclude from these findings that the performance of the agricultural commodity price prediction scheme is dependent not only on the quality of the data utilized for prediction but also on the nature of the algorithm.

Regarding feature importance, the experiment results show that the proposed features are among those with a higher contribution level to the prediction model. The proposed features such as; close, high, low, open, exponential moving average (EMA), double exponential moving average (DEMA), simple moving average (SMA), true-high, true-low, trend, seasonality, and relative strength index (RSI) are among the most important feature in the price prediction scheme.

5.6 Threats to Validity

5.6.1 Threats to Internal Validity

Measurement associated threats, specifically threats to instrumentation might affect the validity of the experiment. Those threats might arise from the type of machine used for running the experiments and the choice of development tools. In this study, however, all approaches are evaluated on the same machine. Hence, the threats will not affect the results and comparisons made in this study.

5.6.2 Threats to External Validity

Threats to external validity might arise from:

- Choice of parameters (sampling rate, selection of feature size, and others): this threat might affect experiments conducted with proposed approach as well as reproduced baseline works. In this study, all approaches are subjected to similar parameter setting (only applicable for common parameters such as sampling rate, selection of feature size and others depending on each approach implementation detail). Hence, threats related to parameter choice will not affect comparisons.
- Selection in commodity type: variety of commodity used for experimentation might also impose another threat to external validity. In this study, however, all approaches are evaluated with commodity types composed of two types of commodities : coffee and sesame.

Chapter 6

Conclusion And Recommendation

6.1 Conclusion

Agricultural commodity price prediction helps the government, investors, and farmers to make informed decisions. Realizing the benefit, several researchers proposed different prediction models that use different features. In this study, we proposed agricultural commodities i.e coffee and sesame, price prediction approach using technical indicators and time series features.

To assess the effect of the seasonality and trend pattern (periodicity), momentum, volatility, trend, volume, and linearity features on agricultural commodity price prediction, we experimented using models built with four machine learning algorithms: Artificial Neural Network, Support Vector Machine, Extreme Learning Machine, and Random Forest. The experiment compares the performance of the baseline approach with the proposed approach considering both seasonality and other commodity properties.

The findings demonstrate that the proposed approach enhances agricultural commodity price prediction performance in all situations except MAE of both coffee and sesame in the Extreme Learning Machine. The RMSE of price prediction is reduced by an average of 4.3, 4.4, 2.7, and 5.1 while using Artificial Neural Network, Extreme Learning Machine, Support Vector Machine and Random Forest, respectively.

From the results of our experiment, we also conclude that considering seasonality and trend (periodicity) properties of commodity and other properties such as volatility, momentum, volume, linearity, and trend, could improve the performance of agricultural commodity price prediction. The features such as close, high, low, open, exponential moving average (EMA), double exponential moving average (DEMA), simple moving average (SMA), true high, truelow, trend, seasonality, relative strength index (RSI) are among the top important features that contributed to the improved agricultural commodity price prediction. Finally, among the four machine-learning algorithms used in this research work, Artificial Neural Network and Extreme Learning Machine performs better towards providing the highest accuracy than the other algorithms in most of the experiment.

6.2 Recommendation and Future works

The results of our study show that considering seasonality and trend (periodicity), momentum, volatility, trend, volume, and linearity properties of agricultural commodity help to improve commodity price prediction. We suggest that further research into additional features could help to improve the performance of the commodity price prediction scheme. In light of this, the following research studies are planned for the future:

- We will do research on forecasting agricultural commodity production using weather data to stabilize commodity prices.
- We will conduct research to reduce high volatility by incorporating certain characteristics influencing the quick increase and fall in commodity prices into the forecast model.
- We suggest investigating the option to build more robust model by applying advanced ensemble techniques

Bibliography

[1] M. M. Mostafa, "Forecasting stock exchange movements using neural networks: Empirical evidence from kuwait," *Expert Syst. Appl.*, vol. 37, no. 9, 6302–6309, 2010, ISSN: 0957-4174.

[2] T. Xiong, C. Li, Y. Bao, Z. Hu, and L. Zhang, "A combination method for interval forecasting of agricultural commodity futures prices," *Knowledge-Based Systems*, vol. 77, Jan. 2015.

[3] M. Yercan and H. Adanacioglu, "An analysis of tomato prices at wholesale level in turkey: An application of sarima model," *Custos e Agronegocio*, vol. 8, pp. 52–75, Nov. 2012.

[4] Y. G. H. Wu, "Prediction of corn price fluctuation based on multiple linear regression analysis model under big data," *Neural Comput & Applic 32, 16843–16855.*, 2020.

[5] V. Jadhav, B. V. Chinnappa Reddy, and G. M. a. Gaddi, "Application of arima model for forecasting agricultural prices," *Journal of Agricultural Science and Technology*, vol. 19, no. 5, 2017.

[6] G. M. Nasira and N. Hemageetha, "Vegetable price prediction using data mining classification technique," *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, pp. 99–102, 2012.

[7] Y. Yu, H. Zhou, and J. Fu, "Research on agricultural product price forecasting model based on improved bp neural network," *Journal of Ambient Intelligence and Humanized Computing*, Aug. 2018.

[8] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[9] G. Y. Dopi, R. Hartanto, and S. Fauziati, "Systematic literature review: Stock price prediction using machine learning and deep learning," in *International Conference on Management, Business, and Technology (ICOMBEST 2021)*, Atlantis Press, 2021, pp. 52–61.

[10] T. W. A. Khairi, R. M. Zaki, and W. A. Mahmood, "Stock price prediction using technical, fundamental and news based approach," *2019 2nd Scientific Conference of Computer Sciences (SCCS)*, pp. 177–181, 2019.

[11] J. S. Vaiz and M. Ramaswami, "Forecasting stock trend using technical indicators with r. forecasting stock trend using technical indicators with r," *International Journal of Computational Intelligence and Informatics, Vol. 6: No. 3,*, 2016.

[12] D. Zhang, S. Chen, L. Liwen, and Q. Xia, "Forecasting agricultural commodity prices using model selection framework with time series features and forecast horizons," *IEEE Access*, vol. 8, pp. 28 197–28 209, 2020.

[13] S. Dametew, "A data analysis and market price prediction of ethiopian commodity market with machine learning algorithms," M.S. thesis, Addis Ababa University, 2018.

[14] Y. Zhang and S. Na, "A novel agricultural commodity price forecasting model based on fuzzy information granulation and mea-svm model," *Mathematical Problems in Engineering*, vol. 2018, pp. 1–10, Nov. 2018.

[15] W. S. L. H. L. L. W. Y. Zhang JD Sun YB, "Forecasting model of agricultural products prices basedon seasonal index adjustment and hgwo-svr algorithm," *Statistics and Decision, Vol. 34No. 19, pp. 33-37.*, 2018.

[16] A. Yeheyis, "Coffee price pridiction using machine-learningtechniques: A case of ethiopian commodityexchange (ecx)," M.S. thesis, Adama Science and Technology University, 2020.

[17] B. P. B V and M Dakshayini, "Performance analysis of the regression and time series predictive models using parallel implementation for agricultural data," *Procedia Computer Science*, vol. 132, pp. 198–207, 2018, International Conference on Computational Intelligence and Data Science, ISSN: 1877-0509.

[18] E. B. D. Bianconcini, "Seasonal adjustment methods and real time trend-cycle estimation.," *Springer.Berlin, Germany*, 2016.

[19] G.-B. Huang, Q.-Y. Zhu, and C. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, Dec. 2006.

[20] H. Li, "Machine learning for the subsurface characterization at core, well, and reservoir scales," 2020.

[21] S. Asgari, "Practical modeling and optimization of ultrasound-assisted bleaching of olive oil using hybrid artificial neural network-genetic algorithm technique," *Computers and Electronics in Agriculture*, vol. 140, Aug. 2017.

[22] M. A. Hall *et al.*, "Correlation-based feature selection for machine learning," 1999.

[23] X.-Q. Zeng and G.-Z. Li, "A supervised solution for redundant feature detection depending on instances," in *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, ser. BIBMW '12, USA: IEEE Computer Society, 2012, 299–306, ISBN: 9781467327466. DOI: 10.1109/BIBMW.2012.6470320.

[24] L. Wang, J. Feng, X. Sui, X. Chu, and W. Mu, "Agricultural product price forecasting methods: Research advances and trend," *British Food Journal*, vol. ahead-of-print, Mar. 2020.

[25] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice.* English, 3rd. Australia: OTexts, 2021.

[26] E. Evans and S. Nalampang, "An analysis of the u.s. demand for avocado (persea americana mill.," *Acta Horticulturae, (831), 247-254. doi: 10.17660/actahortic.2009.831.28*, 2009.

[27] A. Darekar and A. Reddy, "Cotton price forecasting in major producing states," *Economic Affairs, Vol. 62, No. 3, pp. 373-378,*, 2017.

[28] R. Pardhi, R. Singh, and R. Paul, "Price forecasting of mango in varanasi market of uttar pradesh," *Current Agriculture Research Journal*, vol. 6, pp. 218–224, Aug. 2018.

[29] A. Kamu, A. Ahmed, and R. Yusoff, "Forecasting cocoa bean prices using univariate time series models.," *Journal of Arts Science & Commerce*, vol. 1, p. 71, Jan. 2010.

[30] Z. M. Li, S. W. Xu, L. G. Cui, G. Q. Li, X. X. Dong, and J. Z. Wu, "The short-term forecast model of pork price based on cnn-ga," in *Manufacturing Engineering and Technology for Manufacturing Growth*, ser. Advanced Materials Research, vol. 628, Trans Tech Publications Ltd, Feb. 2013, pp. 350–358.

[31] N. Shrivastava and B. Panigrahi, "A hybrid wavelet-elm based short term price forecasting for electricity markets," *International Journal of Electrical Power & Energy Systems*, vol. 55, 41–50, Feb. 2014.

[32] C.-M. Vong, W.-F. Ip, P.-K. Wong, and C.-C. Chiu, "Predicting minority class for suspended particulate matters level by extreme learning machine," *Neurocomput.*, vol. 128, 136–144, 2014, ISSN: 0925-2312.

[33] T. Liu and S. Yin, "An improved particle swarm optimization algorithm used for bp neural network and multimedia course-ware evaluation:" *Multimedia Tools and Applications*, vol. 76, May 2017.

[34] N. Salman, A. Lawi, and S. Syarif, "Artificial neural network backpropagation with particle swarm optimization for crude palm oil price prediction," *Journal of Physics: Conference Series*, vol. 1114, p. 012 088, Nov. 2018.

[35] G. Jiang, M. Luo, K. Bai, and S. Chen, "A precise positioning method for a puncture robot based on a pso-optimized bp neural network algorithm," *Applied Sciences*, vol. 7, p. 969, Sep. 2017.

[36] B. Wang, P. Liu, Z. Chao, *et al.*, "Research on hybrid model of garlic short-term price forecasting based on big data," *Computers, Materials & Continua*, vol. 57, pp. 283–296, Jan. 2018.

[37] D. Zhang, G. Zang, J. Li, K. Ma, and H. Liu, "Prediction of soybean price in china using qr-rbf neural network model," *Computers and Electronics in Agriculture*, vol. 154, pp. 10–17, Nov. 2018.

[38] P. Wang, H. Zhang, Z. Qin, and G. Zhang, "A novel hybrid-garch model based on arima and svm for pm2.5 concentrations forecasting," *Atmospheric Pollution Research*, vol. 8, Feb. 2017.

[39] Q. Duan, L. Zhang, F. Wei, X. Xiao, and L. Wang, "Forecasting model and validation for aquatic product price based on time series ga-svr," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 33, pp. 308–314, Jan. 2017.

[40] C. Luo, Q. Wei, L. Zhou, J. Zhang, and S. Sun, "Prediction of vegetable price based on neural network and genetic algorithm," in *Computer and Computing Technologies in Agriculture IV*, D. Li, Y. Liu, and Y. Chen, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 672–681, ISBN: 978-3-642-18354-6.

[41] X. Ran, Z. Shan, Y. Fang, and C. Lin, "An lstm-based method with attention mechanism for travel time prediction," *Sensors*, vol. 19, p. 861, Feb. 2019.

[42] D.-i. Yoo, "Developing forecasting model of vegetable price based on climate big data," ser. Poster 330-2016-13943, 2015, 2015. DOI: 10.22004/ag.econ.206052.

[43] J. C. Reboredo and M. A. Rivera-Castro, "A wavelet decomposition approach to crude oil price and exchange rate dependence," *Economic Modelling*, vol. 32, pp. 42–57, 2013, ISSN: 0264-9993.

[44] X. Zhang, K. K. Lai, and S.-Y. Wang, "A new approach for crude oil price analysis based on empirical mode decomposition," *Energy Economics*, vol. 30, pp. 905–918, May 2008.

[45] T. Xiong, C. Li, and Y. Bao, "Seasonal forecasting of agricultural commodity price using a hybrid stl and elm method: Evidence from the vegetable market in china," *Neurocomputing*, vol. 275, Dec. 2017.

[46] D. Wang, C. Yue, S. Wei, and J. Lv, "Performance analysis of four decomposition-ensemble models for one-day-ahead agricultural commodity futures price forecasting," *Algorithms*, vol. 10, no. 3, 2017, ISSN: 1999-4893.

[47] F. Azuaje, I. Witten, and F. E, "Witten ih, frank e: Data mining: Practical machine learning tools and techniques," *Biomedical Engineering Online - BIOMED ENG ONLINE*, vol. 5, pp. 1–2, Jan. 2006.

[48] S. Panigrahi and D. H. Behera, "A hybrid ets-ann model for time series forecasting," *Eng. Appl. of AI*, vol. 66, pp. 49–59, Nov. 2017.

[49] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, 1996.

[50] Y. Zhang, S. Wei, L. Zhang, and C. Liu, "Comparing the performance of random forest, svm and their variants for ecg quality assessment combined with nonlinear features," *Journal of Medical and Biological Engineering*, vol. 39, Apr. 2018.

Appendix A

Feature Description and Analysis

The correlation diagram based on mutual information (MI) is shown in Figure A.1. The light point represents the maximum MI value of all the twenty eight features. The greater the correlation, the lighter the color.

Figure A.1 shows the correlation among features before feature reduction. It can be seen that most of the correlations are dark-colored, which reveals that these features contain diverse information on the time series. A few points are light-colored, which implies that the information contained in these features is redundant. These redundancies may have negative effects on the generalization performance of the model. Thus feature reduction should be employed to eliminate the redundancies.

Figure A.2 shows the correlation of the features after feature reduction. Compared to Figure A.1 , the numbers of the dark-colored points have been reduced. This result shows that the CFS approach is a workable approach to feature reduction.

After feature reduction, twenty-five features including, five TS dataset feature, four time-series feature and sixteen technical indicators, remained. In general, the details of the selected feature for price prediction are listed in Table A.1.

TABLE A.1: The reserved features after feature reduction.

Category	Features
Trend	EMA, DEMA, SMA, MACD, ADX
Momentum	CCI, CMO, ROC, RSI, Stoch
Volatility	ATR, TrueHigh, TrueLow
Volume	OBV,CMF, MFI
Periodicity	Seasonality, trend
Linearity	Linearity, nonlinearity
TS dataset	Close, Open, High, Low, volume

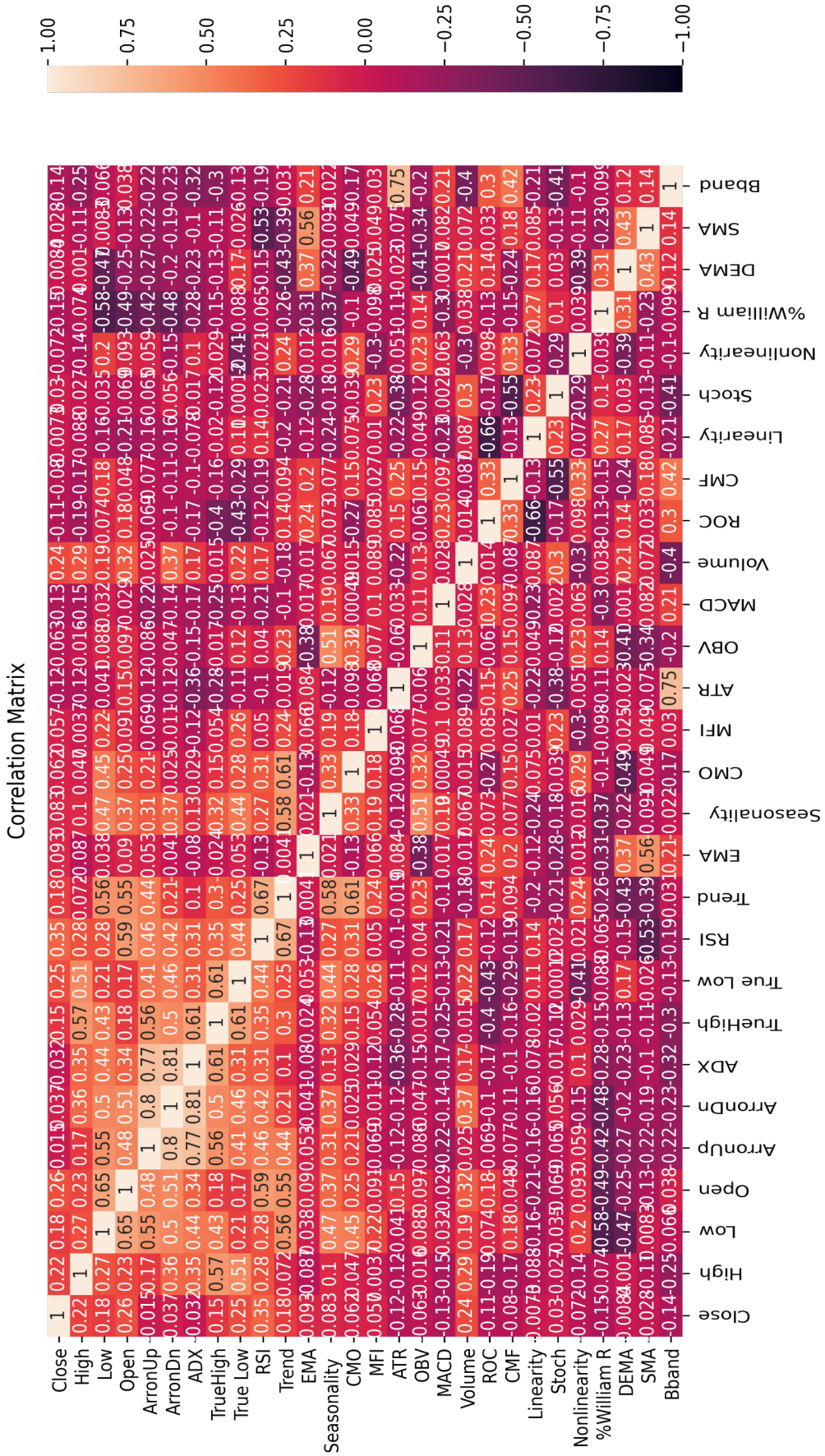

FIGURE A.1: Feature correlation matrix.

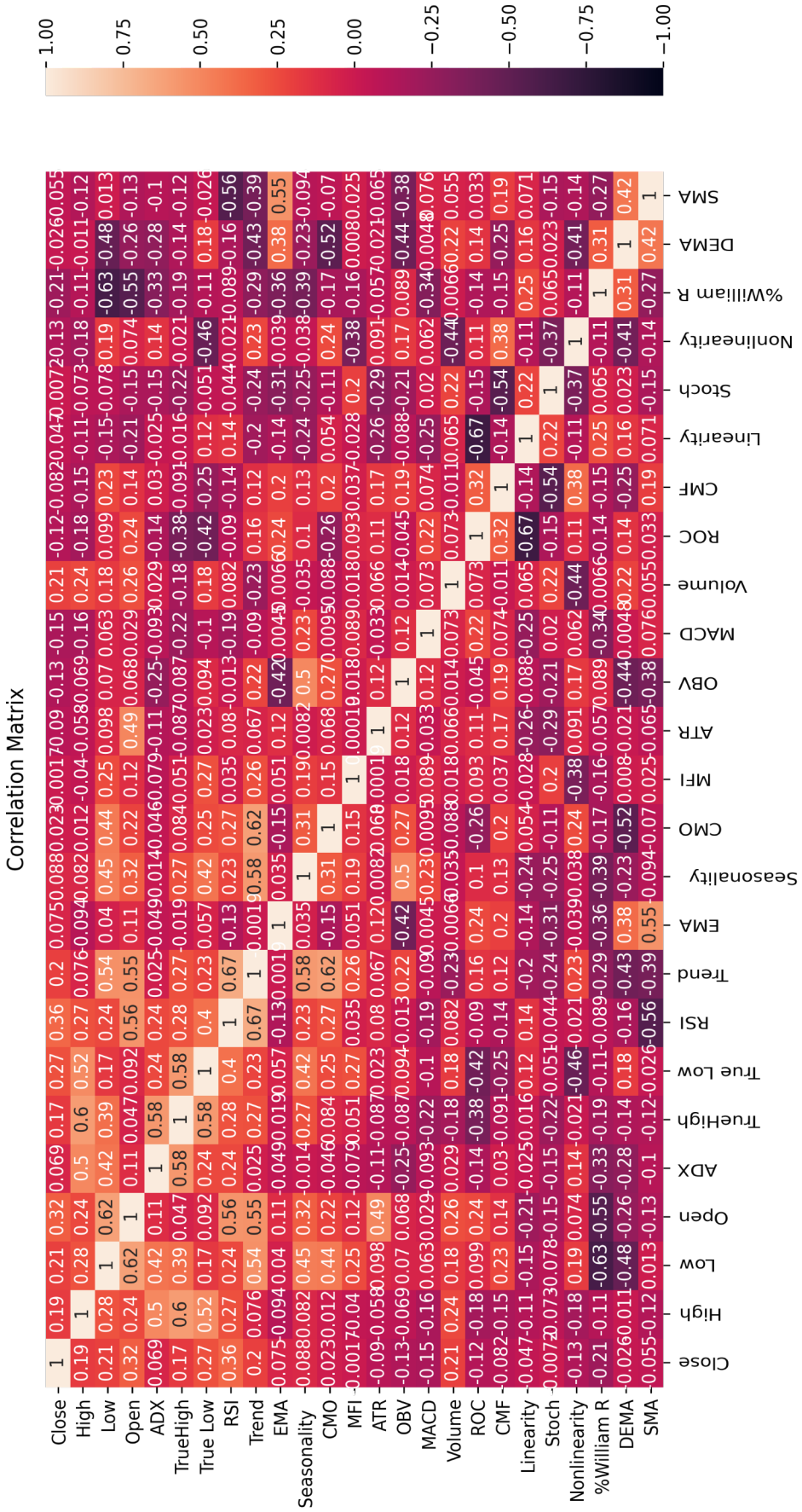

FIGURE A.2: correlation matrix After feature reduction.