A Thesis

Submitted to the School of Electrical and Computer Engineering of

Addis Ababa University

In

Partial Fulfillment of the Requirements For the Degree of Master of Science In

Computer Engineering

A Data Analysis and Market Price Prediction of Ethiopian Commodity Market with Machine Learning Algorithms

By

Selam Damtew

Advisor: Dr Surafel Lemma

March 2018

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

"A Data Analysis and Market Price Prediction of Ethiopian Commodity Market with Machine Learning Algorithms"

By

Selam Damtew

ADDIS ABABA INSTITUTE OF TECHNOLOGY

APPROVAL BY BOARD OF EXAMINERS

Chairman Department of Graduate Committee	Signature
Advisor	Signature
Examiner	Signature
External Examiner	Signature

Declaration

I, Selam Damtew, hereby confirm that the work presented in this thesis is original and my own. Where information has been derived from other sources, I confirm that it has been indicated in the thesis.

Acknowledgment

I would like to dedicate this thesis to my late mother **Yematawork Tiruneh** and late father **Damtew Guche**. I am deeply grateful for my family valued support throughout this entire process and for my husband who provided me with unwavering love and support through this Master's program from start to end.

I would like to express my gratitude to my advisor, **Dr. Surafel Lemma**, whose continual supervision and direction made this paper possible. Thank you for pushing me to engage critically, thoughtfully and to write with clarity of purpose during this iterative exercise.

I also would like to thank ECX and its staffs for their full cooperation for the provision of the required data for my study.

Last but not least I would like to express my gratefulness to my friends and colleagues for supporting me throughout my study. Thank you all!

Abstract

The current Ethiopian market is conducted in a traditional manner and market drivers are still not used for prediction of future market price. Although, large amount of market data have been gathered throughout years by both governmental and non-governmental organizations, yet little have been done to analyze the data for future market price prediction. Moreover, the analysis methods were often manual creating inefficiency in time and quality of market prediction. Analyzing valuable data will show us what the future holds and accelerate the development goals of the country in the sector. The study examines features of current Ethiopian market attributes to find out most valuable features for predicting market price. Eighteen technical indicators are taken and tested for their individual ability of prediction and redundancy. From the feature selection of commodity marke, we have found that features like Stochastic %K, Stochastic %D, Close gain/loss, High, close price, Opening Price, Low, RSI, Ton and Moving Average Convergence/ divergence (MACD) founded to be in the top ten of individual performance evaluation. Moreover features namely Stochastic %K, Relative Strength Index (RSI), Bollinger Bands-Upper, Highest-High, close gain/loss, Simple Moving Average (SMA), Closing price, MACD-Fast, Exponential Moving Average (EMA), MACD-Slow and Low founded to be less redundant. The study also compares four machine learning models for their prediction ability of Ethiopian commodity market price. The outcomes of feature selection were used to compare the models. Two experiments were conducted; the first was comparison of the models with 10 fold cross validation using feature of high individual predictive ability and less redundancy. The second one was a comparison of models with separate train and test data using features of high individual predictive ability and less redundancy. From the models (Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (K-NN) and Ensemble Learning) the performance of ANN and Ensemble Learning algorithms are shown to be accurate than SVM and K-NN. The average MAE rate of the ANN model was 2.8084. Ensemble Learning and SVM follow with average MAE rate of 4.9362 and 8.1178 respectively. The other model was least performer with the MAE rate above 45.3381.

Keywords: Feature selection; Technical Indicators Price prediction; Machine Learning Algorithms

Table of Contents

List of Tables

List of figures

List of Abbreviations

ANN Artificial Neural Network

ADL Accumulation Distribution Line

BF Best First

BPN Back-Propagation Neural Network

CC Correlation Coefficient

CCI Commodity Chanel Index

CFS Correlation- Based Feature Selection

DAX Deustcher Aktienindex

DS Directional Symmetry

ECX Ethiopian commodity Exchange

EMA Exponential moving Average

EPS Earning Per-Share

FN False Negative

FP False Positive

FTSE Financial Time Serious Exchange

GI General Index

IG Information Gain

IPSO Improved Particle Swarm Optimization

K-NN K-Nearest Neighbor

LDA Linear Discriminant Analysis

MACD Moving Average Convergence/Divergence

MAE Mean Absolute Error

MFM Money Flow Multiplier

MFV Money Flow Volume

NAV Net Asset Value

NMSE Normal Mean Squired Error

QUD Quadratic Discriminant Analysis

RAE Root Absolute Error

RMSE Root Mean Squared Error

RSI Relative Strength Index

SVM Support Vector Machine

TN True Negative

TP True positive

TSIR Term Structure of Interest Rates

STIR Short-Term Interest Rate

LTIR Short-Term Interest Rate

CPI Consumer Price Index

IP Industrial Production

GC Government Consumer

PC Private Consumption

GDP Gross Domestic Product

1 Introduction

Market prediction using different analysis techniques is regularly practiced in modern marketing systems by collecting and analyzing different market information [1]. Traders in any part of the world are interested in a market that is profitable and uses different Technical indicators, macroeconomic factors and stock market indexes to study the market. These numerous market drivers information which reflects the existing market price characteristics and facilitates prediction of future market price characteristics [2]. As a result we can prevent anticipated negative changes on the market due to new information of market. However, market analysis is not a common practice in Ethiopia and is often carried out using traditional tools and manual practices making the process time taking and prone to errors. Although other countries' carried out different studies on market prediction, a direct implementation of their findings is not practical. Each study follows different approaches based on the countries economical and market situation. Moreover, the market features that have impact on one country may not have similar impact in another country. As a result we need to take a closer look of the target market to form or amend market strategies.

There are several approaches that are commonly used to predict market price for a given market namely fundamental, technical and quantitative analysis. The fundamental approach examines the economic factors that drive the price of a market [2]. The technical analysis helps to anticipate what others are thinking based on the attribute of goods [3]. In technical analysis indicators are computed from past prices and volumes. The computed values are then used to predict future changes in prices. Technical analysis also identifies regularities by extracting patterns from noisy data and inspecting the goods charts visually. The third, quantitative approach, is more quantitative and statistical which enables to easily program the rules into a computer or machines. Such analysis uses techniques such as statistical arbitrage, automated trading and electronic market making. Quantitative analyses are more efficient in removing the human element of trading and replace it with statistical market models [3].

This thesis explores the predictability of future Ethiopian market price enabling governmental and non-governmental organizations in the sectors, and individual trades to conduct market

activities with minimum business risks. Every day market activities can be conducted with a clear clue of the market prices. Predicting future market price condition enables to decide what product needs to be produced, when and how to address the market [2]. In the context of Ethiopia, to our knowledge market data has not been analyzed in automated manner and no structured market conceptual framework exists. As a result traders are forced to take a huge business risk and are scared to invest because of business uncertainty. Recently, Ethiopian commodity exchange (ECX) started hosting commodity market. Every day ECX disseminates market information on coffee, beans, sesame and grains in real-time bases and offering contracts for further delivery. Although, such systems can be appreciated to certain extent sufficient data analytical activities are not employed to predict future market price scenarios resulting in market uncertainty. Traders are still seeking market analysis which indicate future opportunities and reduce business risk. To explore the information in accurate way we can encode the data in to technical indicators and be classified using known classification techniques. Learning from other countries experience using the advantage of computational algorithms is one way to solve the problem. Computational algorithms can be used to find patterns in data. The output of computational algorithms then can be used to predict future market price of goods in Ethiopian. The study aims to identify market features which influence the prediction of Ethiopian market price. Besides it explores different computational algorithms that are more efficient in predicting market price in Ethiopia

1.1 Problem statement

The current Ethiopian market is conducted in a traditional manner and market data are still not used for prediction of future market price. Although, large amount of market data have been gathered throughout the years by both governmental and non-governmental organizations, yet little have been done to analyze the data and use it for future market price prediction. Traders start business without appropriate current and future market information. Even in governmental sectors the data collected on goods price is just left as it is without further treatment or analysis for future development or action. The analyses performed so far are in small size and manually, which is time taking and prone to human errors. In most cases the approach followed fails to do the work effectively and are affected by different human factors.

In this study, we particularly investigate,

- Which Ethiopian market features are most useful in predicting future market price?
- We selected four machine learning models namely: SVM, ANN, K-NN and Ensemble Learning. Which model is more accurate for predicting future market price of Ethiopian?

1.2 Objective

1.2.1 General Objective

The general objective of this study is to analyze existing market data and predict Ethiopian market price for key marketed commodities. The data gathered from different governmental and non-governmental sectors will be analyzed to come up with a more accurate prediction of Ethiopian market price.

1.2.2 Specific Objective

- Identify most valuable market features for predicting future market price in Ethiopia.
- From the selected machine learning models we identify which model better suits the market situation and prediction of Ethiopian market price.

1.3 Scope

ECX host a large market of commodities and collects data from the market activities on a daily bases. Such detailed and organized data empower us to know the past better and predict future price with greater degree of certainty. However, market price prediction cannot cover all market sectors of Ethiopian due to its broadness. This study focuses only on goods that ECX collects namely Coffee, Sesame and Pea bean.

1.4 Thesis Outline

The following sections include description of background in Chapter 2; a literature review of related works in Chapter 3; a proposed methodology followed for the study in Chapter 4; result and evaluation in Chapter 5; conclusion of the study in Chapter 6 and recommendation in Chapter 7.

2 Background

2.1 Market Features

There are a number of market features including macroeconomic inputs, Stock market index and technical indicators that are used for market price prediction and can be used separately or in combination. For this chapter we will discuss what these features indicate and a more detailed research review will be presented in chapter three.

2.1.1 Definition of Macroeconomic inputs terms

Term Structure of Interest Rates (TS) [4]: is the relationship between interest rate or bond yields and different maturities.

Short-term Interest Rate (ST) [4]: the rate of short term borrowings which will be affected between financial institutions.

Long-term Interest Rate (LT) [4]: refers to governmental bonds maturing in 10 years. The rate is determined by the price charged by the lender, the risk from the borrower and the fall in capital value.

Consumer Price Index (CPI) [5]: measures changes in the prices of goods and services that households consume. Such changes affect the real purchasing power of consumers 'incomes and their welfare.

Industrial Production (IP) [6]: is a principal short-term economic business indicator, which aims to measure at a monthly frequency the ups and downs of industrial production during the long period of time.

Government Consumption (GC) [7]: is the difference between a country's exports and its imports of goods, services, and income. It measures the difference between a country's national income and its domestic expenditure on consumption and investment.

Private Consumption (PC) [8]: is the value of the consumption goods and services acquired and consumed by households.

Gross Domestic Product (GDP) [9]: measures the market value of all final goods and services produced *within a country* in a given period of time.

2.1.2 Definition of Stock Market index terms

Dow Jones [10]: a price-weighted average of 30 significant stocks traded on the NY Stock Exchange and Nasdaq. It includes companies like General Electric, Disney and Microsoft.

NASDAQ [11]: Stock Market is an American stock exchange and is the second-largest stock exchange in the world by market capitalization. The NASDAQ Index has more than 3000 components representing the performance of technology and growth companies.

S&P CNX NIFTY [12]: a stock index endorsed by standard & poor and composed of 50 of the largest and most liquid stocks found on the National Stock Exchange of India's. Commonly they are used to represent the market for benchmarking India's.

BANC NIFTY [13]: represent the 12 most liquid and large capitalized stocks from the banking sectors which trade on the National Stock Exchange. It provides investors and market intermediaries a benchmark that captures the capital market performance of Indian banking sectors.

CNX INFRA [14]: includes companies belonging to Telecom, power, roads, railways and other utility services provider. It can be used for a verity of purpose such as benchmarking fund portfolios, launching of index funds and structural products.

CNX 100 [15]: it is India's stock market index. It is diversified 10 stock indexes accounting for 38 sectors of the economy.

Hang Seng [16]: it is a weighted index of the 40 largest companies that trade on the Hong kong exchange. The indexes capture the leadership of Hong Kong exchange, and covers approximately 65% of its total market capitalization. It is classified into four main lines of business including commerce and industry, finance, utilities and properties.

DAX [17]: it consists 30 major German companies trading on the Frankfurt stock Exchange.

Nikkei [18]: it represents Japans stock market. It is a price-weighted index comprised of Japans top 225 blue-chip companies on the Tokyo stock exchange.

FTSE [19]: share index of 100 companies listed on the London Stock Exchange with the highest market capitalization.

2.1.3 Definition of Technical Indicators

In the study of market movements technical indicators create a way to represent price and volume data to simplify discovering of decision making points. The following are common technical indicators used in the study.

Moving Average [21]: is a type of finite impulse response filter used to analyze a set of data points by creating a series of averages of different subsets of the full data set. Given a series of numbers and a fixed subset size, the moving average can be obtained by first taking the average of the first subset. The fixed subset size is then shifted forward, creating a new subset of numbers, which is averaged. This process is repeated over the entire data series.

Moving Average shows what direction a currency pair is going and where potential levels of support and resistance. Moving averages are constructed by finding the average closing price of a currency pair at any given time and then plotting these points on a price chart. It can be divided in to simple and exponential moving average.

A simple moving average (SMA) is calculated by computing the average price over a specified number of periods. This process is repeated for each day, forming its own time series as shown in equation 1.

$$SMA = \frac{Pc(t) + Pc(t-1) + \cdots + Pc(t-n-1)}{n} \quad \textbf{(1)}$$

Exponential moving average (EMA) is calculated by taking a weighted average of past prices as described in equation 2. More recent values of the security affect the EMA result more than past values.

$$EMA = K\big(Pc - EMA(t-1)\big) + EMA(t-1) \quad \textbf{(2)}$$

Where

$$k = \frac{2}{1+n} \quad \textbf{(3)}$$

n is a period of EMA and Pc is the closing price of a current day. The first value of EMA is calculated from SMA.

Moving Average Convergence/Divergence [20]: is a technical analysis indicator created by Gerald Appel in the late 1970s. It is used to spot changes in the strength, direction, momentum, and duration of a trend in price. The MACD is a collection of three signals that are calculated using exponential moving averages. First, the MACD which is the difference between 12 days and 26 days exponential moving average as described in equation 4. The second collection is the signal which is 9 days EMA of the MACD as shown in equation 5. Finally there is the histogram, which is the difference between the MACD and the signal as shown in equation 6. It is also referred to as Price Oscillator when the MACD has different values for the moving averages. It is used to identify changes in the strength, direction and momentum of a security.

$$MACD = EMA(12) - EMA(26) \quad \textbf{(4)}$$

$$signal = EMA(MACD, 9) \quad \textbf{(5)}$$

$$Histogram = MACD - signal \quad \textbf{(6)}$$

Relative Strength Index [20]: is a technical indicator used in the technical analysis of financial markets. It is intended to chart the current and historical strength or weakness of a market based on the closing prices of a recent trading period. It is plotted on a scale of 0-100 which makes it easy to quantify and measure trading signals the indicator generates. RSI is used to measure the velocity and direction of price movements.

$$RSI = 100 - \left(\frac{100}{1 + \frac{U}{D}}\right) \quad \textbf{(7)}$$

Where U- average value of the positive price changes over n days;

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

Bollinger Bands [23]: are trending indicators that can show you not only what direction a currency pair is going but also how volatile the price movement of the currency pair is. Bollinger Bands are bands relating to volatility placed below and above a moving average. Bollinger Bands consist of a simple moving average, an upper Bollinger Band and a lower Bollinger Band. The simple moving average is used to compute a moving average as described in equation 8.

$$MA = SMA(n) \quad (8)$$

The upper Bollinger Bands are sum of the moving average and that of standard deviation (σ) of closing price (*Pc)* over the past *n* days with a selected multiple of K as shown in equation 9.

$$BBupper = MA + K\sigma \quad (9)$$

The lower Bollinger Bands are the difference between moving average and that of standard deviation (σ) of closing price (*Pc)* over the past *n* days with a selected multiple of K as shown in equation 10.

$$BBlower = MA - K\sigma \quad (10)$$

Stochastic Oscillator [24]: is a momentum indicator that uses support and resistance levels. The Stochastic Oscillator is a momentum indicator that refers to the current price in relation to its price range over a period of time. It is composed of two lines, the %K and the %D as shown in equation 11 and equation 12 respectively. These lines represent predicted turning points for the price of a security.

$$\%K = 100\left(\frac{\left(Pc - Pl(n)\right)}{ph(n) - Pl(n)}\right) \quad (11)$$

$$\%D = EMA(\%K, 3) \quad (12)$$

Where *Pc* is the closing price of the current day and *Pl* and *Ph* are the low and high prices over the past period n.

Commodity Channel Index [25]: The commodity channel index (CCI) is an oscillating indicator that can show how bullish or bearish traders are toward a currency pair and how dramatic those sentiments are. CCI measures a security variation from its statistical mean which

is used to identify cyclical trends. It is calculated by taking the difference between the current price of a security and it's SMA, divided by the mean absolute deviation of the price as illustrated in Equation 13.

$$CCI = \frac{Pt - SMA(Pt)}{(0.015)\sigma mad(Pt)} \quad (13)$$

Where

$$Pt = \frac{Ph+Pl+Pc}{3} \quad (14)$$

σmad is the mean absolute deviation of price of period t (Pt).

P_h high price

P_l low price

P_c close price

Accumulation/distribution index [21]: is a technical analysis indicator intended to relate price and volume in the stock market.

$$CLV = \frac{(close - low) - (high - close)}{high - low} \quad (15)$$

This ranges from -1 when the close is the low of the day, to +1 when it's the high. For instance if the close is 3/4 the way up the range then CLV is +0.5. The accumulation/distribution index adds up volume multiplied by the CLV factor i.e.

$$\text{accdist} = \text{accdist}_{\text{prev}} + \text{volume} * \text{clv} \quad (16)$$

The starting point for the acc/dist total, i.e. the zero point, is arbitrary, only the shape of the resulting indicator is used, not the actual level of the total.

Chaikin Oscillator [26]: it is defined as the difference between the 3 day EMA of the Accumulation Distribution Line and the 10-day EMA of the Accumulation Distribution Line and used to relate price and volume data for a particular security. Like other momentum indicators it

designs to anticipate directional changes in the Accumulation Distribution Line by measuring the momentum behind the movements. Chaikin oscillator can be given by

$$chaikin\_oscillator = EMA(ADL, 3) \quad (17)$$

Where,

$$Accumulation\_Distribution - Line(ADL) = ADL(t-1) + MFV \quad (18)$$

$$Money\_Flow\_Multiplier(MFM) = \frac{(Pc - Pl) - (Ph - Pc)}{Ph - Pl} \quad (19)$$

$$Money\_Flow\_volume(MFV) = MFM * v \quad (20)$$

Momentum and Rate of Change [27]: The rate of change is a simple technical indicator calculated by taking the difference between today's closing price and the price of the same security n days ago. This is then scaled by the older closing price as presented in equation 22

$$Momentum = Pc - Pc(t-n) \quad (21)$$

$$ROC = \frac{Pc - Pc(t-n)}{Pc(t-n)} \quad (22)$$

Volatility [20]: most frequently refers to the standard deviation of the continuously compounded returns of a financial instrument within a specific time horizon. Volatility measures variation of price over a period of time as presented in equation 23. Can be used to confirm price behavior and expanding and contracting ranges highlight the strength of breakouts and trends. It is calculated by taking the standard deviation of n closing prices divided by a simple moving average of n days.

$$Volatility = \frac{\sigma(Pc, n)}{SMA(Pc, n)} \quad (23)$$

Where Pc closing price, SMA simple moving average and n period

Although all the features are discussed, our study only focuses on the use the technical indicators, as macroeconomic input data is not available and stock markets are not yet introduced in Ethiopia.

2.2 Feature selection Algorithms

Feature selection plays an important role in data analysis process by extracting relevant and non-redundant features. It refers to choosing a subset of the original input variables. The selected features represent the original dataset characteristics better and prediction with these features can improve accuracy [28].Selecting key features of market have a greater impact on predicting market price. To this end, we plan to use feature selection techniques.

2.2.1.1 Correlation-based Feature Selection (CFS)

CFS [28] is employed to handle feature redundancy. The feature selector is simple and fast to execute. It eliminates irrelevant and redundant data and, in many cases, improves the performance of 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 method measures the correlation between the attribute and the class, with the hypothesis that an ideal set of features should be highly correlated with the class, yet uncorrelated with other features. This is to ensure that redundancies and numbers of features are minimized. This hypothesis is built from two aspects: the usefulness of individual features for prediction and the level of inter-correlation among them. It can be stated as:

$$\text{merits} = \frac{\text{K}r_{\text{cf}}}{\sqrt{\text{K} + \text{K}(\text{K}-1)r_{\text{ff}}}} \quad (24)$$

Where $Merit_s$ is the "merit" of a feature subset S containing k features;

$$r_{ff} = \sum_{fi \in S} \frac{1}{K} \sum (fi, c) \quad (25)$$

Is the mean feature-class correlation ($f \in S$, and c is the class), an indication to how easily a class could be predicted based on the feature; and r_{ff} is the average feature inter-correlation 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:

$$IG(X,Y) = H(X|Y), \tag{26}$$

$$H(X) = -\sum_{i} P(x_i) \log_2\big(P(x_i)\big), \tag{27}$$

$$H(X|Y) = -\sum_{j} P(y_j) \sum_{i} P(x_i|y_j) \log_2\big(P\big(x_i|y_j\big)\big) \tag{28}$$

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, CFS 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 selects the most relevant features and by the greatest extent avoids the re-introduction of redundancy. As other filter models, CFS does not need to reserve any training data for the subsequent evaluation. Besides, it works well on smaller data sets.

2.3 Machine Learning Algorithms

There are many machine learning algorithms that can be used to classify a problem given a set of features. Machine learning algorithms have been explored and used for high frequency trading and market microstructure data [38]. These work looks to those algorithms to see if any are particularly useful in classifying Ethiopian commodity market data and give price prediction

given a set of inputs generated through technical analysis. Algorithms investigated are: Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Artificial Neural Network (ANN) Ensemble learning (see **Chapter 5**).Below we have briefly described the selected algorithms.

2.3.1 Support Vector Machine (SVM)

Support Vector Machines are supervised learning models that can be used for classification, prediction and clustering problems. An SVM takes a set of input observations and associated binary outputs 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; linearly separating the observation sets [29]. The margin around this linear separation is calculated to be as large as possible. A partitioning in higher dimensional space by a linear hyper-plane corresponds to a nonlinear partition in the output space [30]. This higher dimensional partitioning is known as the SVM kernel, and can be defined by any mathematical surface. Some of the more common kernels are linear, quadratic, polynomial and Gaussian radial basis function.

Having a set of observation an SVM tries to linearly separate the data using a hyper-plane. In the case of two dimensions a simple line can be drawn to separate the data. As we can see from **Fig 1** the bold line linearly separates the data. The distance from the middle line to side lines shows the margins.

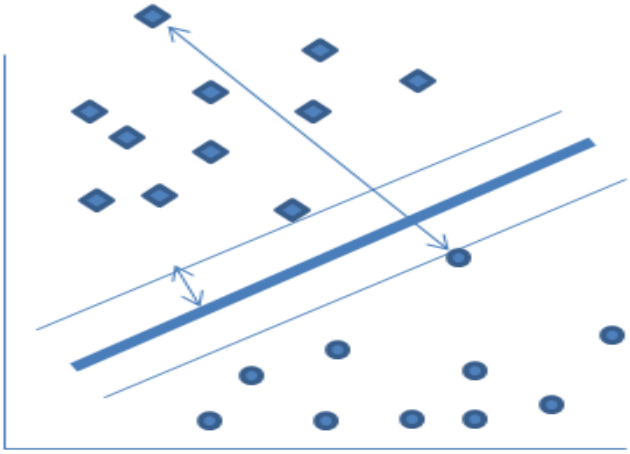

Figure 1 A linear line separating the data types

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 that it founds far from the middle line is the most accurate prediction.

In the case of real world application, it is not usually possible to get a line that perfectly separates the data within the space. Hence, we might have to use a curved decision boundary. It is possible to get a hyper-plane which could separate the data but this may not be desirable if the data has noise in it. In such cases we need to use the soft margin method [29]. The soft margin method allows for points to appear on the incorrect side of the margin. These points have a penalty associated with them. The penalty increases as the points are farther from the margin. The hyper-plane separation looks to minimize the penalty of incorrectly labeled points, while maximizing the distance between the remaining examples and the margin.

The other approach employed to separate data that isn't linearly separable is to map the data into a higher dimensional space using an SVM kernel. By mapping $x=(x, x^2)$ the data will be mapped into two dimensional space. When this two dimensional mapping is graphed, an obvious linearly separable line appears. The mapping used to increase the dimensionality of the problem is dependent on the data space being investigated. The above computations, which are used to find the maximum-margin separator, can be expressed in terms of scalar products between pairs of data points (in the high-dimensional feature space). These scalar products are the only part of the computation that depends on the dimensionality of the high-dimensional space. So if we had a fast way to do the scalar products we would not have to pay a price for solving the learning problem in the high-dimensional space. The kernel trick is just a magic way of doing scalar products a whole lot faster than is usually possible. It relies on choosing a way of mapping to the high-dimensional feature space that allows fast scalar products [30].

2.3.2 K-nearest Neighbor

The K-Nearest Neighbor (KNN) classifier is one of the simplest machine learning algorithms [3]. The purpose of KNN algorithm is to use a database in which the data points are separated into

several separable classes to predict the classification of a new sample point. An object is classified by looking to its nearest examples. The measurement can be performed using any distance metric and a majority vote [42]. K states how many neighbors will be used in voting. K=1 simply states that an object will be assigned the same class as its nearest example. As the number of K increases then we need to classify a given instant based on the resemblance of all the stated K instances.

In general, we start with a set of data, each data point of which is in a known class. Then, we will be able to predict the class of a new data point based on the known classifications of the observations in the dataset. The process of choosing the class of the new observation is known as the classification problem. Before we can decide whether two observations are similar, we need to find some way of comparing objects. The trouble with this is that our data could be of many different types - a number, a color, a geographical location, a true/false (boolean) answer to a question, etc - which would all require different ways of measuring similarity. The first step is preprocessing the data in the database in such a way as to ensure that we can compare observations. Then, our observations become points in space and we can interpret the distance between them as their similarity (using some appropriate metric). The other problem comes here, how to decide which observations from the database are similar enough to our new observation for us to take their classification into account when classifying the new observation [41]. One of the most widely used metrics is the *Euclidean distance*. The Euclidian distance between two instances($X_1, X_2, X_{3,\ldots}X_n$) and ($U_1, U_2, U_{3,\ldots}U_n$) is given by the following formula

$$\sqrt{(X_1 - U_1) + (X_2 - U_2) + \cdots + (X_n - U_n)} \quad \textbf{(29)}$$

Where X_1, X_2, X_3, X_n are predictors of the instance #1 and U_1, U_2, U_3, U_n are predictor of the instance #2.

2.3.3 Artificial Neural Networks

An artificial neural network (ANN) is an interconnected group of nodes intended to represent the network of neurons in brain [39]. 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), an input layer represented as $x_1 \ldots, x_6$, an optional hidden layer, and an output layer y. The goal of the ANN is to determine a set of weights w (between the input, hidden, and output nodes) that minimize the total sum of squared errors. During training these weights w_i are adjusted according to a learning parameter $\lambda \in [0, 1]$ until the outputs become consistent with the output. Large values of λ may make changes to the weights that are too drastic, while values that are too small may require more iterations (called epochs) before the model sufficiently learns from the training data.

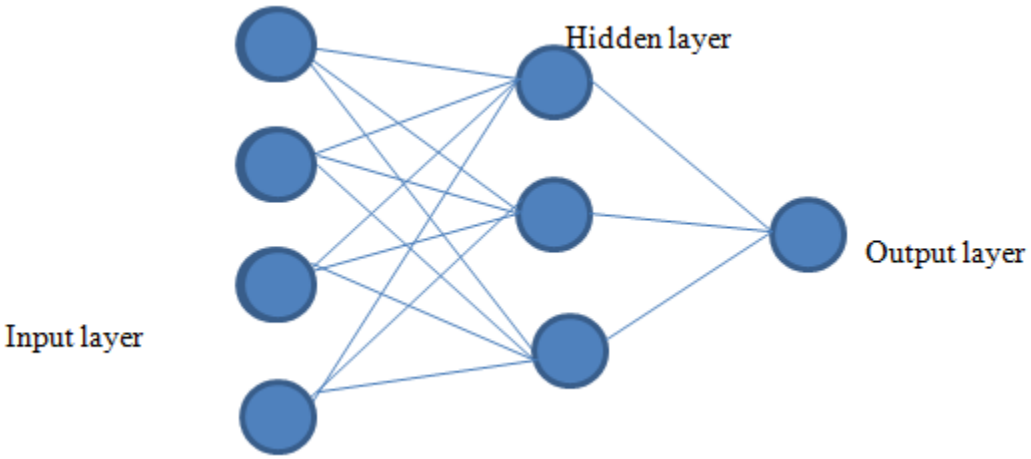

Figure 2 Simple artificial neural networks

The difficulty of using artificial neural networks is finding parameters that learn from training data without over fitting (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 falls below some predetermined error level [39].

2.3.4 Ensemble Learning

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted vote of their classification [45]. The original ensemble

method is Bayesian averaging but more recent algorithms include error-correcting output coding, bagging and boosting.

Because uncorrelated errors of individual classifiers can be eliminated through averaging ensembles are often much more accurate than the individual classifiers that make them up. There are different methods of constructing ensembles namely [45]; (1) Manipulating the Training Examples, (2) Manipulating the Input Features, (3) Manipulating the Output Targets and (4) Injecting Randomness.

The first method manipulates the training example to generate multiple hypotheses [46]. The learning algorithm is run several times each time with a different subset of the training examples. This technique works especially well for unstable learning algorithms whose output classifier undergoes major changes in response to small changes in the training data. The second method manipulates the set of input features available to the learning algorithm [45]. The method usually only works when the input features are highly redundant. The third method manipulates the y values that are given to the learning algorithm [46]. Having a large class, K, new learning problems can be constructed by randomly portioning the K classes into two subsets A and B and give a level of 0 and 1 respectively for the subsets. The relabeled data from the subsets then given to learning algorithm which then can construct the classifier. The last method injects randomness into the learning algorithm [45].

3 Literature Review

Researches have been working on market data analysis and market price prediction using the advantage of Machine Learning models. The following is an overview of the papers that served as guides to this work. The overview covers the four selected models (SVM, ANN, K-NN and Ensemble Learning) and in addition it looks for other hybrid models, genetic Algorithms and other algorithms that have been used to market prediction.

3.1 Support Vector Machine

Kim uses an approach similar to KNN trading system method, replacing KNN with a support vector machine for classification [30]. From the Korea composite stock price index, 12 technical indicators are generated to be used as input variables. Using a Gaussian radial basis function for the SVM's kernel, Kim explores which parameters perform best for the stock data. A comparison between the SVM classifier, a back propagation Neural Network (BPN) and a KNN is performed. SVM is found to be sensitive to the value of its parameters and SVM was able to outperform the BPN and KNN classifiers in experimental tests when the correct parameters were selected.

Prasad and Padhy [32] explore two machine learning techniques: Back Propagation Technique (BP) and Support Vector Machine (SVM) for predicting futures trade prices in Indian stock market. The study uses the data collected by Stock Exchange (NSE) of India. A number of stock market indexes namely S&P CNX NIFTY, BANC NIFTY, S&P CNX 500, CNX INFRA and CNX 100 are used as features for predicting future price. The performance metrics used are the normalized mean squared error (NMSE), mean absolute error (MAE) and directional symmetry (DS). The NSME for all the futures stock index taken into consideration fall in the range of 0.9299 to 1.1521. The MAE fall in the range of 0.2379 to 0.3887 and the last cnidarian DS starts from 55.17 to 91.2512. The results show that SVM performs better than BPN.

Shunronget.al, [41] tried to develop a new prediction algorithm that exploits the temporal correlation among global stock markets and various financial products to predict the next-day stock trend with the aid of SVM. They consider different stock market indexes around the globe as an input features for next-day stock prediction. They have examined each index for their ability of predicting next-day stock market. For single feature prediction they found DAX which is Germany's stock index more accurate than the other stock indexes with accuracy of 70.8%. In long term prediction they conclude that prediction accuracy increases when time span becomes longer. Regarding multi-feature prediction they compare SVM and MART which is a decision tree based boosting algorithm for 4 selected features and all available features. The result shows that SVM performs best for 4 selected features with accuracy of 74.4% and MART performs best using all the available features (73.9%). Finally they develop a new model which takes the output from the prediction of SVM and takes it to make marketing rules and decision.

3.2 K-Nearest Neighbor (K-NN)

Jeffrey and Bushee [1] tried to develop a market trading model that can successfully trade market securities for a profit, beating buy-and-hold. In the study, 12 technical indicators, 54 features and 10 macroeconomic data indicators were constructed for classifying daily stock market data. The data used ware collected from yahoo finance which covers a time period from January 1, 2001 to January 1, 2010. This system uses the K-NN, SVM classifiers for prediction. The results showed the strategy based on KNN model outperformed the buy-and-hold strategy for 7 of the 10 stocks. K-NN also outperformed the SVM model in prediction.

Subha et.al [3] studied the predictability of stock index movement of the popular Indian Stock Market indices. The study uses two Indian market indices namely BSE-SENSEX and NSE-IFTY. The prediction model considers opening value, high value, low value and closing value of the market index as independent variables and the next day's closing value as the dependent variable. The results of k-NN classifier were compared with the Logistic regression model and it was observed that the k-NN classifier outperforms the traditional logistic regression method as it classifies the future movement of the BSE-SENSEX and NSE-NIFTY more accurately.

3.3 Artificial Neural Network (ANN)

Zabir et.al [33] uses ANNs to forecast Bangladesh Stock Exchange market index values with reasonable degree of accuracy. They used back propagation algorithm for training session and Multilayer feed-forward network as a network model for predicting price. General Index (GI), Net Asset Value (NAV), P/E ratio, Earnings per Share (EPS) and Share Volume ware found to be inputs that affect the share price. By using the past historical data of ACI pharmaceutical company which include only 2 inputs, they predict stock values for future 8 days of November 2010 from back-Propagation algorithm and were able to compare the predicted values with the real values. The average error of the simulation was 3.71 percent. They used past historical data of ACI pharmaceutical company which include only 5 inputs, and predict stock values for future 8 days of November. The average error of the simulation was 1.53 %. From this they conclude that the more input data we have the better training and more accurate results.

Adebiyi.et.al [34] focused on improving the accuracy of stock price prediction by using the hybrid approach. The hybrid model combines the variables of technical and fundamental analysis. Then created neural network predictive model for stock price prediction. The study used three-layer (one hidden layer) multilayer perceptron models (a feed-forward neural network model) trained with back-propagation algorithm. Historical stock prices of different companies were obtained from published stock data available in the site. The hybridized approach takes identified 18 input variables to train the network comprising both technical variables and fundamental analysis variables. Training data and testing data were selected and they observed the various outcomes of the different network structure models implemented with Matlab Neural Network Tool Box version 7. For the implementation of their hybridized approach, they experimented with different neural network model configurations. The most accurate daily stock price prediction was 18-24-1 back-propagation network (BPN) using the hybridized approach. The hybrid approach combines the variables of technical and fundamental analysis. Finally they conclude that the hybridized approach has the potential to enhance the quality of decision making of investors in the stock market by offering more accurate stock prediction compared to existing technical analysis based approach.

3.4 Ensemble Learning

Narayanan and Govindarajan [44] analyzed the time based data set and forecast the stock market price more precisely than the existing models. The paper used, classifier techniques namely Support Vector machines and Naïve Bayes commonly for analyzing time based data sets. They also used ensembles model which expands the classification accuracy by creating more than one classifier. The accuracy of the given algorithm was improved by boosting method. Boosting is a process of joining weak or inappropriate prediction rules for creating a machine based learning method to increase the prediction accuracy and to minimize the error rate. The algorithm pertains weights on the training set. At the starting point, the weights are evenly distributed. But for the subsequent training, the weights are increased for the not properly classified example.

The Historical Time Series data were collected from www.datamarket.com. The record attributes ware placed as Date, cname, Start, Max, Min, End, and Quantity for any point of day. The dataset was separated for training and testing. The training dataset consists of 3777 records and testing dataset consists of 2500 records.

To measure the performance of the model they computed accuracy and classification error. Accuracy is closeness to a measured value or the standard set. In time series analysis, the forecasting value which is nearest to the actual value is taken as accuracy. The formula for accuracy is $A= (TP+TN)/(TP+FP+FN+TN)$ in which TP indicates the True Positive, TN indicates True Negative, FN indicates False Negative and FP indicates the false Positive. The classification Error (E) of any technique, **t‟** is the cases not correctly classified ($FP+FN$). The formula for calculating classification Error is $Et= F*N*100$ where t represents the technique, F denotes number of items classified incorrectly and N reveals total number of samples

The result showed that the proposed AdaBoost Support Vector Machine and Naïve Bayes are superior to individual approaches for stock market prediction problem in terms of classification accuracy and error. The accuracy and classification error of SVM is 93.86% and 6.14%

respectively where as Naïve Bayes is 88.32% and 11.68% respectively. For the same set of input data, the proposed AdaSVM produces 94.33% of accuracy and 5.67% of classification error whereas the proposed AdaNaive produces 97.19% of accuracy and 2.81% of classification error.

3.5 Hybrid Models

Tsai et.al [31] created a model which combines Artificial Neural Network (ANN) and decision trees to enhance the rate of prediction accuracy in stock price forecasting model. Fundamental and technical analyses are used as indicators for hybrid model which forecasts the stock price in electron industry in Taiwan. The dataset collected from the TEJ database and 53 Variables ware selected. The results showed that the performance scored by the hybrid model outperforms the performance of individual models.

Wei et.al [2] investigated the predictability of financial movement direction with SVM by forecasting the weekly movement direction of NIKKEI 225 index. Besides they propose a combining model by integrating SVM with the other classification methods namely Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Elman back-propagation Neural-Networks. Macroeconomic inputs, Stock market Index and exchange rate are considered as input features. Historical data is collected from Yahoo finance and Pacific Exchange rate. The results show that based on the individual performance SVM takes the lead. However, the highest performance is achieved by the combined model of the four approaches.

3.6 Genetic Algorithm

The research done by Karazmodeh et.al [42] used Improved Particle Swarm Optimization (IPSO) based on Support Vector Machines. They were able to efficiently predict stock indices. PSO is an evolutionary computation technique which works similarly as birds travel when trying to find sources of food, or similarly the way a fish school will behave. The behavior was modeled that the "particles" inside the "swarm" (or population) are treated as solutions to a given

problem. Having this algorithm they combined it with SVM and come up with the hybrid model IPSOSVM. Technical indicators and stock market index ware used as a feature. From the results they conclude that mutation in particles led the accuracy being higher, since the particle always searches the whole state space which prevents the mistakes of not finding the other best options in other possible states. The hybrid model is found more effective than the individual models (PSO and SVM).

3.7 Other Algorithms

Soni and Shrivastava [43] explored classification and regression tree (CART), linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). The models ware used for classification of Indian stock market data. And give simple interpretation of stock market data in the form of binary tree, linear surface and quadratic surface respectively. Data has been taken from stock market trading company situated at India. Comparisons of machine learning algorithms have been done on the basis of the misclassification and correct classification rate. They observed that classification and regression tree machine learning classifier performance was better than linear and quadratic discriminant analysis classifier in the misclassification rate and correct classification rate.

Although many studies have explored the area of market movement and price prediction, a direct implementation of these studies output is not practical for Ethiopian market. The market characteristics differ from country to country. Moreover most of the studies address stock market which is established in Ethiopian. Although recently the country is organizing commodity market so this study aims to explore predictability of agricultural commodity market price.

4 Proposed Methodology

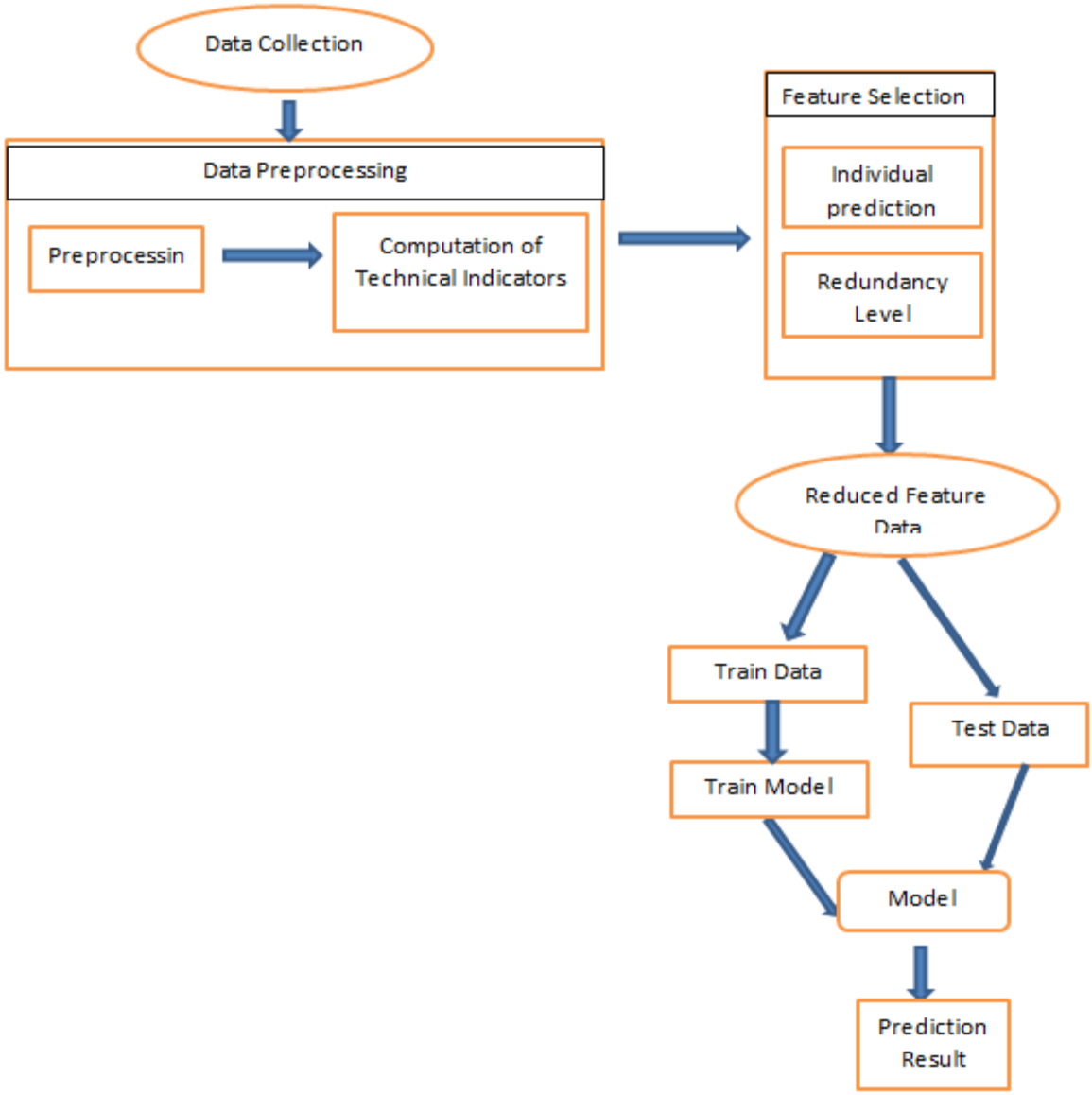

Figure 3 proposed methodology

Fig 3 showed the direction of the proposed methodology. The proposed methodology starts with data collection and the data was preprocessed and technical indicators were computed and included in to the original dataset from ECX. The newly formed dataset was used as an input for feature selection. The features were selected based on their high predictive abilities of individual features and for lower level of redundancy. Then the selected features were used to train and test the four machine learning models. Lastly, the models were compared and best performing predictive machine learning models were used to predict commodity prices.

4.1 Data collection

All the data for this work was collected from Ethiopian Commodity Exchange (ECX).The collected data covered the time period of January 2008 up to January 2015. The data from ECX contained attributes namely; Trade date; lowest price (low), highest price (high), volume (ton), opening and closing price of pea bean, coffee and sesame. To assure the relevance of the dataset, the data attributes were compared with different standard datasets and stock market indexes used for prediction. To overcome the effect of the data gap that comes from the closing of market in weekdays, interpolation of the missing values was used. Interpolation is a process of including and processing externally fetched data in to dataset. The process estimates the value of a function at a point from its values at nearby points.

4.2 Method of data quality assurance

The datasets were checked for completeness and correctness of the required attributes and integrity before analysis and prediction. Details of data exploration techniques and interpolation are discussed in the immediate sections below.

4.2.1 Checking the dataset contains the required attributes.

The study explored different studies related to market price prediction using technical indicators. Moreover we have investigated how to compute different technical indicators (see **Chapter 2**). As a result we now know the required attribute (column names) for computation of technical indicators for our study. Our data from ECX contained required attributes namely; trade date, Opening price, Closing Price, High, Low and volume. The attributes may found in difference name and addition of some other attributes depending on the market they are related. However the data we have contains enough attributes to compute a required indicators that later can be considered as features of the market.

4.2.2 Conducting data integrity test

In this study, we considered a data to be missing if the values for the weekdays are not recorded. Trade Date column were used t locate the missing days from the dataset and missing values were observed in the three of original data sets (coffee, sesame and pea bean) whose values estimated using interpolation considering the neighbors and the overall values that specific date across all years in the data set and the estimated values were then put in a separate column. Accordingly the two data frames were created containing the original data set and full dates of the year respectively. For each trading day at least a single record was taken while multiple records were also included.

4.3 Information about the experiment data

All the data for these work is collected from Ethiopian Commodity Exchange (ECX). The data from ECX contains daily opening and closing price of pea bean, coffee and sesame. Historical price of the three commodities from 2008 up to 2016 is used for the experiment. The data contained 94,993 rows in which the majority of the records are of coffee, which were around 72,160. Sesame and pea beans have 18,021 and 4,812 rows respectively. Form the data we can understand that the coffee trade is conducted throughout the year but the other two crops are seasonal and the trade of these commodities takes place only in few months of the year. The data from ECX contains 6 columns and in addition to those attributes the computed attributes are included. The attributes explored for the study are described in **Table 1.**

Table 1 the initial 24 potential features that may be used for the feature selection.

No	Attribute	Description
1	Trade Date	Date of trade
2	Closing Price	The final price the commodity is sold with
3	High	The highest price given by bidders
4	Low	The lowest price given by bidders
5	Ton	Volume of the commodity provided for bid
6	Opining Price	Opining price stated by the bidder
7	EMA	Exponential Moving Average

8	Close Gain/Loss	The Gain or Loss from the previous day market
9	RSI	Relative Strength Index
10	SMA_20	Simple Moving Average of 20 days
11	SMA	Simple Moving Average
12	BB-Upper	Upper Bollinger Bands
13	BB-Lower	Lower Bollinger Bands
14	MACD Fast	Fast Moving Average Convergence/Divergence
15	MACD Slow	Slow Moving Average Convergence/Divergence
16	MACD	Moving Average Convergence/Divergence
17	MACD Signal	Moving Average Convergence/Divergence
18	Highest High	The highest high over the look up period
19	Lowest Low	The lowest low over the look up period
20	Stochastic %K	Calculated with other quantity %D
21	Stochastic %D	Sample moving average of %K
22	20-Days Mean Deviation	20 days mean deviation

4.4 Features Selection

From a given set of data attributes of Ethiopian market we computed Technical Indicators that latter can be taken as features and from these features we selected the better one which represents the market.

4.4.1 Individual Features Ability

The first thing we do is to see individual features ability in prediction using models such as ReliefFAttributeEval.The ReliefFAttributeEval (see **Chapter 4**) method evaluates individual predicting ability from the given subset of attributes.

4.4.2 Non Redundant Attributes

Another approach is to handle or remove features redundancy with models like Correlation-Based Feature Selection (CSF) and Cfssubsetevalmethods (see **Chapter 2** & **Chapter 4**). The two models were used to select out the redundant attributes that have similar information. Based on the results from the methods, redundant attributes were removed.

4.5 Selection of Machine Learning Models

The principal goal of this research is to analyze existing market data of Ethiopia and predict market price using computational algorithms. With this in mind, the study aim was to discover which machine learning technique is better in predicting future Ethiopian market opportunity and price. Throughout this work different machine learning algorithms were explored and effective ones were used to find patterns in the data. The performance of each machine learning models was tested using valuable features extracted from the dataset in feature selection stage. Then based on successfulness in making predictions and stability in their performance best performing algorithms were selected. The machine learning techniques selected for this thesis were SVM, K-NN, ANN and Ensemble Learning. Each was selected based on their advantages and past performance seen in other research.

Support vector machine is selected due to the following reasons; (1)Data classification could be performed without making strong assumptions; (2) SVM is established on the structural risk minimization principle, which seeks to minimize an upper bound of generalization error, and is shown to be very resistant to the over-fitting problem [11, 12] and (3) SVM model is a linearly constrained quadratic program so that the solution of SVM is always globally optimal, while other models may tend to fall into a local optimal solution [11, 12].

ANN is included in these work because; (1) A neural network can be used to solve linear as well as non-linear programming tasks [6], (2) As a component of an ANN fails, the net continues to operate (based on its highly parallel nature) [6], (3) A neural network *learns* and does not have to be re-programmed [4] and (4) An ANN can be used to solve *classification*, *clustering*, and *regression* related problems [6]

K-nearest Neighbor is selected because; (1) The cost of learning process is zero, (2) Learning does not require making any assumption about the characteristics of the concepts [8] and (3) Complex concepts can be learned by local approximation using simple procedures [8].

Ensemble Learning is included in these work because [21]; (1) Ensemble learning is combine predictions from multiple models so the results are more diversified and (2) More robust estimate of a statistical quantity with a low bias and a high variance.

4.6 Result Comparison

4.6.1 10 fold cross validation

For training and testing we used 10 fold cross validation method provided in weka. This means that the dataset is split into 10 parts, the first 9 are used to train the algorithm, and the 10th is used to assess the algorithm. This process was repeated giving each of the 10 parts of the split dataset a chance to be the held-out test set.

4.6.2 Separate training and test set

Separate training and testing set was prepare and used for the study. The test set will contain a data of different year that is not used for training.

4.7 Performance Metrics

To see the applicability and performance of the above Machine learning techniques different metrics will be used. The effectiveness of classification algorithms may depend on a number of factors like quality of information the attributes provide, the class distribution of the dataset and the number of instances. Such factors were addressed in feature selection stage and have less impact on the performance of the machine learning techniques. The following performance metrics are provided by Weka 3.8.0 and were used for measuring the performance of machine learning techniques.

Correlation coefficient: measures how strong a relationship is between two variables.

Mean absolute error: measure the average magnitude of the errors in a set of prediction, without considering the direction. It expresses average model prediction error in units of the variable of interest.

Root Mean Squared error: is a quadratic scoring rule that also measures the average magnitude of the error. It is square root of the average of squared differences between predicted and actual value.

Relative Absolute error: 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.

4.8 Tool

For the experiment purpose we plan to use weka 3.8.0 which is a data mining tool. The tool is enriched with different classifiers, clustering and attributes selection methods. For feature individual predictive ability evaluation purpose we will use ReliefFAttributeEval and for redundancy check we use CfsSubsetEval. Following are a brief description of the weka functions.

4.8.1 ReliefFAttributeEval:

Evaluates the worth of an attribute by repeatedly sampling instance and considering the value of the given attribute for the nearest instance of the same and different class. Can operate on both discrete and continuous class data.

OPTIONS

numNeighbours-- Number of nearest neighbors for attribute estimation.

sampleSize-- Number of instances to sample. Default (-1) indicates that all instances will be used for attribute estimation.

seed-- Random seed for sampling instances.

sigma-- Set influence of nearest neighbors. Used in an exp function to control how quickly weights decrease for more distant instances. Use in conjunction with weight By Distance. Sensible values = 1/5 to 1/10 the number of nearest neighbors.

weightByDistance-- Weight nearest neighbors by their distance.

4.8.2 CfsSubsetEval :

Evaluates the worth of a subset attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred.

OPTIONS

locallyPredictive-- Identify locally predictive attributes. Iteratively adds attributes with the highest correlation with the class as long as there is not already an attribute in the subset that has a higher correlation with the attribute in question.

missingSeperate-- Treat missing as a separate value. Otherwise, counts for missing values are distributed across other values in proportion to their frequency.

For the purpose of machine learning models we use weka classification available in weka Explorer GUI. There are a seven categories of classifier from we have used functions, lazy and tree. The functions category provides as with The ANN and SVM models. The lazy contains K-NN model. The last one contains the Ensemble Learning model.

5 Result and Evaluation

5.1 Research Question

Market prices of goods in Ethiopia have been collected formally and informally by different authorities. An accumulated data of market can give us most of the information to formulate and direct future market and enables us to see different market opportunities. The traditional way of analyzing market is time taking and labor intensive. In most cases this approach fails to do the work effectively and affected by human factors.

To address the indicated problems, we proposed to encode the data into technical indicators and use Machine learning algorithms (see **Chapter 4**). To see if the proposed methodology holds, we conducted an investigation using the Ethiopian commodity market data. In particular, we investigated the following research questions (RQ).

> RQ 1 [Feature Selection] Does every market features have equal significance in predicting Ethiopian commodity market?

In this RQ, we investigated if every market feature has significance in predicting Ethiopian commodity market. The two activities listed below were done separately and they are not sequential.

- Checking for individual feature prediction ability
- Redundancy within the features

> RQ 2 [Machine Learning comparison] which machine learning algorithms gives a better prediction?

In this RQ, we investigated if every machine learning algorithms has equal performance in prediction future market price. Using the results from feature selection we check for the performance of some selected prediction algorithm.

5.2 Feature selection (RQ1)

5.2.1 Individual performance

The first activity associated with Feature selection is to testing each individual attributes contribution on the predicted prices. The attribute selection method used to see the performance of each attribute was ReliefFAttributeEvalwhich is available in weka 3.8.0 tool (see **Chapter 4**).

Table 2 Features in order of importance, from higher to lower

	ReliefFAttributeEval attribute evaluator
Pea Bean	
	Attributes: 20,21,8,2,3,6,4,9,5,16,14,22,7,11,10,15,12,13,18,19,17,1 %K, %D, Close Gain/Loss, Closing Price, High, Opining Price, Low, RSI, Ton, MACD, MACD Fast, 20-Days Mean Deviation, EMA, SMA, SMA_20, MACD Slow, BB-Upper, BB-Lower, Highest High, Lowest Low, MACD Signal, Trade Date
Sesame	
	Attributes: 20,21,8,2,3,6,4,9,5,16,14,22,7,11,10,15,12,13,18,19,17,1 %K, %D, Close Gain/Loss, Closing Price, High, Opining Price, Low, RSI, Ton, MACD, MACD Fast, 20-Days Mean Deviation, EMA, SMA, SMA_20, MACD Slow, BB-Upper, BB-Lower, Highest High, Lowest Low, MACD Signal, Trade Date
Coffee	
	Attributes: 20,21,8,3,2,6,4,9,5,16,14,22,7,10,11,15,12,17,18,19,13,1 %K, %D, Closing Price, Close Gain/Loss, High, Opining Price, Low, RSI, Ton, MACD, MACD Fast, 20-Days Mean Deviation, EMA, SMA_20, SMA, MACD Slow, BB-Upper, MACD Signal, Highest High, Lowest Low, BB-Lower, Trade Date

The result from ReliefFAttributeEval attribute evaluator includes a number of attributes. The given 22 attributes from **Table 1** were ranked based on their predictive ability. For the three datasets (pea bean, sesame and coffee) used in the study a relatively similar results were recorded (**Table 2**). Attributes namely; %K, %D, Close gain/loss, High, close price, Opening Price, Low, RSI, Ton and MACD were founded in top 10 in their predictive ability with the respective order.

5.2.2 Redundancy removal

The other activity in feature selection is removing redundancy from the given attribute set using CFs subset evaluation and correlation attribute selector of weka 3.8.0 (see **Chapter 4**).The results is presented in the table below.

Table 3 Features in order of redundancy, from lower to higher

	Cfssubsetevalua	Correlation attributeEvaluation
Pea Bean		
	5 Attribute High Close Gain/Lose RSI MACD %K	Attributes: 20,9,8,18,16,11,12,6,4,15,13,14,19,7,3,17,21,1,2,5,10,22 %K, RSI, Close Gain/Loss, Highest High , MACD, SMA, Closing Price, BB-Upper, Low, MACD Slow , BB-Lowe, MACD Fast, Lowest Low, EMA, High , MACD Signal, %D, Trade Date, Opining Price, Ton, SMA_20, 20-Days Mean Deviation
Sesame		
	3 Attributes Ton BB-lower MADC Signal	Attributes: 18,16,12,20,9,8,11,6,4,15,13,14, 19,21,7,2,17,1,3, 5,10,22, Highest High, MACD, BB-Upper, %K, RSI, Close Gain/Loss, SMA, Low, MACD Slow, BB-Lower, MACD Fast, Lowest Low, %D, EMA, Opining Price , MACD Signal,

		Trade Date, High , Ton, SMA_20, 20-Days Mean Deviation
Coffee		
	3 Attributes High RSI MACD Fast	Attributes: 20, 11,4,9,8,18,15,3,2, 17,12,14,19,7, 13,5,6,16,21,1,10,22 %K, SMA, Low, RSI, Close Gain/Loss, highest-high , MACD Sow, High, Closing Price, MACD Signal, BB-Upper , MACD Fast, Lowest Low, EMA, BB-Lower, Ton, Opining Price , MACD, %D, Trade Date, SMA_20, 20-Days Mean Deviation

The result from correlation attribute selection included 22 attributes which were ranked based on their redundancy level in ascending order (**Table 3**). Then 10 least redundant attributes which are commonly found in the three data sets were selected for prediction. Attributes namely; %K, RSI, BB-Upper, Highest-High, close gain/loss, SMA, Closing price, MACD-Fast, EMA, MACD-Slow and Low were found commonly in all data sets. Contrastingly, the CfssubsetEval attribute selector resulted different less redundant attributes for the three commodities. Attributes namely, High, Close Gain/Lose, MACD, %K, Ton, BB-Upper, MADC Signal, RSI, and MACD Fast were found less redundant (**Table 3**). This indicates that, if individual results from commodities are taken for prediction they may not be enough for the prediction as the information will be limited and these will affect prediction negatively.

The results from this two feature selection activities contained 5 attributes in common. Features namely; %K, RSI, Closing Price, Close gain/loss and low are found highly predictive and less redundant. The top ten features results from the two feature selection activities will be used for comparison of machine learning models. We limit the number of features to 10 because of the computational cost and processing time needed for prediction using the entire feature is high. The first group contains top ten features from individual predictive ability and the second group contains the top ten features which were found less redundant.

5.3 Machine Learning Models Comparison (RQ2)

The experiments are conducted using Weka 3.8.0. The four Machine learning models (SVM, ANN, K-NN and ensemble Learning) were used to predict the data on the three selected commodities (coffee, sesame, pee bean). For modeling the machine learning algorithms we used 10 fold cross validation and a separate training and test set. These evaluations are done using the features from feature selection stage.

5.3.1 10 fold cross validation

Group 1: the first group contains top ten attributes from results of individual prediction ability. (%K, %D, Close gain/loss, High, close price, Opening Price, Low, RSI, Ton and MACD)

Table 4 comparison of machine learning algorithm with highly predictive features (10 fold cross validation)

Machine Learning models	Performance matrix	Coffee	Sesame	Pea Bean	Average MAE
SVM	Correlation coefficient	0.9989	0.9997	1	
	Mean absolute error (MAE)	7.4722	10.7132	1.3197	6.5017
	Root mean squired error	11.4213	17.1327	3.399	
	Relative absolute error	1.1123	1.8337	0.379	
	Root relative squired error	2.3219	2.5259	0.8186	
ANN	Correlation coefficient	0.9999	0.9997	0.999	
	Mean absolute error	1.7435	13.6394	3.2007	6.1945
	Root mean squired error	6..1304	20.6303	5.9318	
	Relative absolute error	0.599	1.9527	0.9193	
	Root relative squired error	1.6773	2.5609	1.4288	
K_NN	Correlation coefficient	0.9943	0.9991	0.9959	
	Mean absolute error	11.6119	23.6833	27.2145	20.8365
	Root mean squired error	15.3371	34.4151	37.5646	
	Relative absolute error	10.0975	3.3907	7.8164	
	Root relative squired error	10.6206	4.2716	9.0485	
Ensemble Learning	Correlation coefficient	0.9999	0.9998	0.999	
	Mean absolute error	3.2789	10.5438	3.6502	5.8243
	Root mean squired error	5.6727	16.9661	5.5164	
	Relative absolute error	1.1264	1.5095	1.0484	
	Root relative squired error	1.5621	2.1058	1.3288	

The 10 fold cross validation result indicated that Ensemble Learning prediction recorded the lowest MAE Value of 5.8243, followed by ANN (6.1945) and SVM (6.5017). Contrastingly, machine learning algorithm K-NN recorded an extremely higher MAE (20.8365) indicating the least predictive ability of the model to predict the price of the studied Ethiopian market commodities (**Table 4**). The results for the Ensemble Learning prediction was consistent with all the three commodities and also showed a moderated MAE compared to the other models while MAE values were not consistent across the commodities. Pea bean price prediction using SVM recorded the smallest MAE of 1.3197 followed by MAE value 1.7435 for coffee using ANN. The prediction for commodity sesame has recorded the highest MAE for all prediction models except for K_NN (**Table 4**).

Group 2: the group includes comparing prediction models using the selected top ten attributes from redundancy check. (%K, RSI, BB-Upper, Highest-High, close gain/loss, SMA, Closing price, MACD-Fast, EMA, MACD-Slow, Low)

Table 5 comparison of machine learning algorithm with less redundant features (10 fold cross validation)

Machine Learning models	Performance matrix	Coffee	Sesame	Pea Bean	Average MAE
SVM	Correlation coefficient	0.9997	0.9995	0.9999	
	Mean absolute error (MAE)	11.9754	14.6144	*2.925*	9.8382
	Root mean squired error	18.9214	21.8231	*4.8099*	
	Relative absolute error	2.0031	2.4298	0.8309	
	Root relative squired error	2.9934	4.0152	1.1586	
ANN	Correlation coefficient	0.9999	0.9998	0.9997	
	Mean absolute error	1.7435	12.6824	4.1777	6.2012
	Root mean squired error	6..1304	17.4741	10.2386	
	Relative absolute error	0.599	1.8257	1.1999	
	Root relative squired error	1.6773	2.1689	2.4687	
K-NN	Correlation coefficient	0.996	0.999	0.9982	
	Mean absolute error	23.4699	23.6833	16.8145	21.1125
	Root mean squired error	32.8255	34.4151	25.5616	
	Relative absolute error	8.0627	3.3907	4.8104	
	Root relative squired error	8.9813	4.2716	6.1815	
Ensemble Learning	Correlation coefficient	0.9999	0.9997	0.9998	
	Mean absolute error	3.4266	12.0638	5.169	6.8864
	Root mean squired error	7.5895	19.1424	8.8332	
	Relative absolute error	1.1771	1.7275	1.4846	
	Root relative squired error	2.0665	2.378	2.1288	

From Table 5 we computed the average MAE for the four models across the three commodities. We found that the average MAE to be 9.3383 for SVM, 6.2012for ANN, 21.1125 for K-NN and 6.8864 for Ensemble Learning. Based on the average value ANN takes the first place. Ensemble Learning prediction showed a closer result to ANN model, while SVM and K-NN showed relatively increased average MAE rate. The smallest MAE (1.7435) was recorded for ANN prediction model in coffee which is similar to the result obtained using top ten predictive features (**Table 4**) followed by MAE of 2.925 recorded for pea bean using SVM.

Figure 4 showed that the performance of the top 10 features from individual feature selection has exceeded over the non-redundant features for the SVM and Ensemble Learning. For the model K-NN the difference was marginal and for the case of ANN it was found insignificant. On average we can say that the features from Individual feature selection has superiority over the non-redundant features.

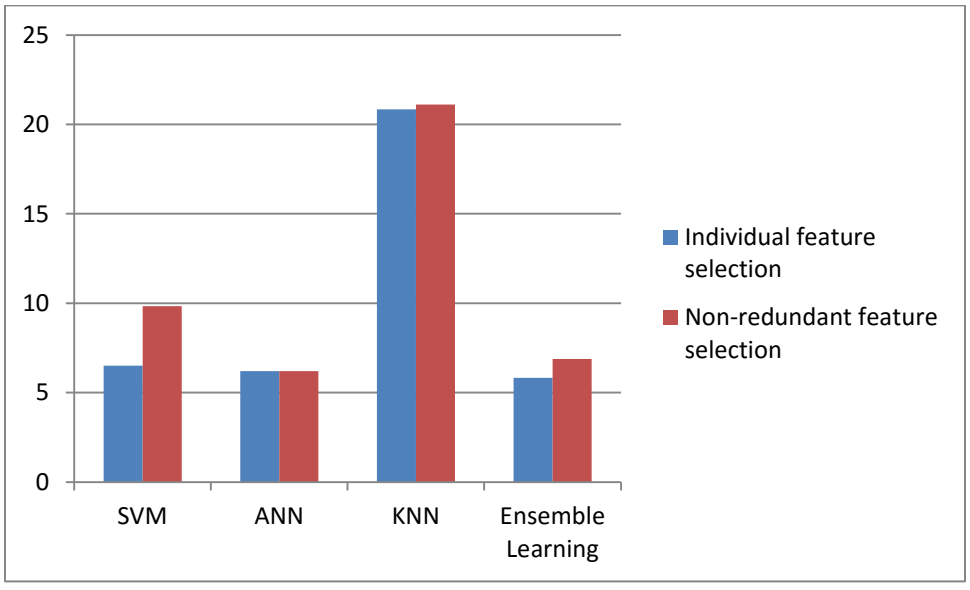

Figure 4 Comparison of individual and non-redundant feature selection for 10 folds cross validation

5.3.1.1 Experiments with Separate Training and Test data

In this section separate train and test set were prepared for the experiment to test the prediction models. The test set contained a data of different year that was not used for training. Prediction model comparison for coffee and pea bean were made using 84 % of the data for training and the other 16% for testing purpose, while the proportion for sesame was 74% for training and 26% for testing. The proportion differ for the three data set because of the difference in the amount of the collected data.

Group 1: the first group contained top ten attributes from results of individual prediction ability. (%K, %D, Close gain/loss, High, close price, Opening Price, Low, RSI, Ton and MACD)

Table 6 comparison of machine learning algorithm with highly predictive features (separate train and test data)

Machine Learning models	Performance matrix	Coffee	Sesame	Pea Bean	Average MAE
SVM	Correlation coefficient	0.9873	0.9992	0.9998	
	Mean absolute error (MAE)	11.5981	9.9888	2.7667	8.1178
	Root mean squired error	18.5323	14.2259	5.2227	
	Relative absolute error	3.1473	1.5739	0.354	
	Root relative squired error	4.1032	2.176	0.6429	
ANN	Correlation coefficient	0.9999	0.9998	0.9997	
	Mean absolute error	1.2166	5.6839	1.5248	2.8084
	Root mean squired error	3.7112	8.561	5.7375	
	Relative absolute error	0.4201	0.8626	0.1951	
	Root relative squired error	1.0184	1.1054	0.7063	
KNN	Correlation coefficient	0.9935	0.9946	0.9351	
	Mean absolute error	29.7984	33.7384	72.4777	45.3381
	Root mean squired error	39.2275	47.2856	88.6036	
	Relative absolute error	11.7035	5.1199	9.1146	
	Root relative squired error	11.1156	8.1055	10.7325	
Ensemble Learning	Correlation coefficient	0.9999	0.9997	0.9996	
	Mean absolute error	3.3094	7.2889	4.2103	4.9362
	Root mean squired error	4.9128	10.9571	6.3391	
	Relative absolute error	1.1426	1.1061	0.5295	
	Root relative squired error	1.3481	1.4148	0.7679	

From **Table 6** we computed the average MAE for the four models across the three commodities. We found the average MAE to be 8.1178 for SVM, 2.8084 for ANN, 45.3381 for K-NN and 4.9362 for Ensemble Learning. The results for the ANN prediction was consistent with all the three commodities and also showed a moderate MAE of 2.8084 compared to the other models. Moreover, coffee recorded the smallest MAE value of 1.2166 while the second lowest value was for pea bean (1.5248) using ANN prediction model.

Group 2: the group included top 10 attributes from redundancy check. (%K, RSI, BB-Upper, Highest-High, close gain/loss, SMA, MACD-Fast, EMA, MACD-Slow, Low)

Table 7 comparison of machine learning algorithm with less redundant features (Separate train and test data)

Machine Learning models	Performance matrix	Coffee	Sesame	Pea Bean	Average MAE
SVM	Correlation coefficient	0.9879	0.9989	0.9994	
	Mean absolute error (MAE)	17.5981	14.3827	2.5296	11.5034
	Root mean squired error	23.3926	19.2493	7.7355	
	Relative absolute error	9.4471	2.8739	0.3177	
	Root relative squired error	11.1038	3.5176	0.9358	
ANN	Correlation coefficient	0.9987	0.9994	0.9994	
	Mean absolute error	2.1506	10.7593	2.3581	5.0893
	Root mean squired error	17.3067	15.4856	7.7374	
	Relative absolute error	0.8447	2.059	0.2962	
	Root relative squired error	4.9041	2.3596	0.936	
K_NN	Correlation coefficient	0.983	0.9936	0.9538	
	Mean absolute error	47.4905	30.3003	58.1106	45.3004
	Root mean squired error	64.4414	49.7341	71.9338	
	Relative absolute error	18.6523	5.7986	7.0469	
	Root relative squired error	18.2606	7.5786	8.7008	
Ensemble Learning	Correlation coefficient	0.999	0.9997	0.9976	
	Mean absolute error	6.3659	11.9548	14.473	10.9312
	Root mean squared error	16.5434	18.2373	18.5643	
	Relative absolute error	2.5002	1.7081	1.8177	
	Root relative squired error	4.6878	2.2627	2.2458	

From **Table 7** we computed the average MAE for the four models across the three commodities and found the average MAE to be 11.5034 for SVM, 5.0893 for ANN, 45.3004 for K-NN and 10.9312 for Ensemble Learning. The results for the ANN prediction recorded the lowest MAE of for all commodities studied as well as the overall smallest of MAE (2.1506) for coffee. Moreover the result of coffee and Pea bean which is recorded small was seen for the model ANN. Ensemble learning, SVM and K-NN was ranked respectively.

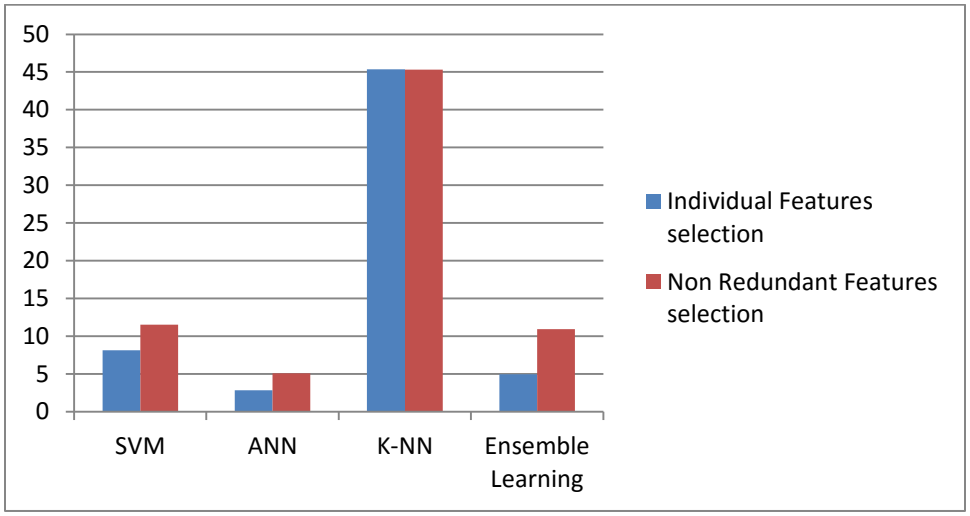

Figure 5 Comparison of individual and non-redundant feature selection for separate train test set

Fig 5 showed that the performance of the top 10 features from individual feature selection has exceeded over the non-redundant features for the SVM, ANN and Ensemble Learning. For the model K-NN the difference was insignificant. On average we can say that the features from Individual feature selection has superiority over the non-redundant features.

In this work the four machine learning models have been analyzed and applied to the commodity price data. The best performance was achieved by ANN, while K-NN model had the worst performance on the commodity market data. Ensemble Learning and SVM show a comparable performance. The models ANN, Ensemble Learning and SVM achieve an average MAE of

5.0733, 7.1445 and 8.9902 respectively. A direct comparison of our work with other literatures may be difficult. Most of the literatures ware done using a combination Stock market Index's, macroeconomic inputs and technical indicators. Our work ware limited to technical indicators due to the unavailability of macroeconomic inputs and un-introduction of Stock market. Besides the prediction of market price was not study in country side. Although the three models namely ANN, SVM and Ensemble Learning have a considerable performance in the area of financial time serious data prediction around the world. This has illustrated in the work of Zabir et.al [6], Kandananond [48], kim [9], Prasad and Padhy [5] and Narayanan and Govindarajan [21].

5.4 Prediction values for ANN and Ensemble Learning

The experiments results showed in the above section separate the three models (ANN, Ensemble learning and SVM) with minimum MAE difference. The analysis of variance showed that these prediction models were significantly different (p=0.05) MAE values. Prediction models (ANN, Ensemble learning and SVM) recorded significantly lower MAE as compared to K_NN indicating that the model has lowest prediction performance for the studied commodities (see **appendix 9.2**). Meanwhile, ANN, Ensemble Learning and SVM can be used for predicting the price of the commodities as the MAE value difference among them was insignificant. In the current study market price for 5 days ahead was predicted using top two models *viz* ANN and Ensemble Learning. The prediction used the highly predictive features as they are showed a better performance when compared to less redundant features (see **Fig 4 and Fig 5**). Following are the price prediction results for the three commodities namely; coffee, sesame and pea bean. The predictive prices for the selected models showed in **Table 8.**

Table 8 Prediction values for ANN and Ensemble Learning

	Days	Actual	Predicted for ANN	Predicted for Ensemble Learning
Pea bean	Day 1	825	824.89	829.39
	Day 2	760	759.91	756.08
	Day 3	885	885.09	880.45

	Day 4	858	848.23	854.42
	Day 5	945	944.99	945.44
Coffee	Day 1	1030	1024	1021
	Day 2	796	795	802
	Day 3	1400	1406	1402
	Day 4	1178	1176	1177
	Day 5	760	756	757
Sesame	Day 1	2700	2676.89	2680.5
	Day 2	2237	2251.83	2275.79
	Day 3	3370	3376.89	3370.45
	Day 4	3320	3314.24	3310.87
	Day 5	2670	2631.93	2645.49

6 Conclusion

The paper raised two research questions and performs the research activity. The first RQ was examining features of current Ethiopian market attributes to find out most valuable features for predicting market price. Features for the study are derived from the collected data. We have computed 18 technical indicators. The computed indicators are taken as a feature. We evaluate for features individual predictive ability and the redundancy level. From the feature selection of commodity market we have found that features like (%K, %D, Close gain/loss, High, close price, Opening Price, Low, RSI, Ton and MACD founded in the top ten of individual performance evaluation. Moreover features namely %K, RSI, BB-Upper, Highest-High, close gain/loss, SMA, Closing price, MACD-Fast, EMA, MACD-Slow and Low founded less redundant from the given dataset. These results are categorized in two groups and used as an input for the machine learning algorithms.

The second research question was comparison of machine learning models that better predict the market price. The outcomes of feature selection were used to compare the models. We conduct two experiments; the first was comparison of the models with 10 fold cross validation using feature of individual predictive ability and less redundancy. The second one was a comparison of models with separate train and test data using feature of individual predictive ability and less redundancy. From the models (SVM, ANN, K-NN and Ensemble Learning) the performance of ANN and Ensemble Learning algorithms were showed superior on SVM and K-NN. The average MAE rate of the ANN was found to be 5.0733. Ensemble Learning and SVM follows with MAE rate 7.1445 and 8.9902 respectively. The K-NN model was least performer with the MAE rate of 33.2964.

7 Recommendation

As discussed in the paper we only explore technical indicators. Interested one can look for other features rather than features derived from technical indicators. The effects of other features derived from microeconomic factors or market indexes are not explored so this can be a good start for them. In addition to this the paper only look for individual performance of the machine learning model, however the performance gain due to a hybrid of the models did not covered. So other researches can be done by hybrid the explored models of this paper. Moreover researchers can also look for some other models that are not explored in this paper.

Reference

[1] Subha, M. V., and S. ThirupparkadalNambi. "Classification of Stock Index movement using k-Nearest Neighbours (k-NN) algorithm."*WSEAS transactions information science and application, Issue* 9 (2012).

[2] Caley, Jeffrey Allan. *A Survey of Systems for Predicting Stock Market Movements, Combining Market Indicators and Machine Learning Classifiers*.Diss. Portland State University, 2013.

[3] Huang, Wei, YoshiteruNakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine."*Computers & Operations Research* 32.10 (2005): 2513-2522

[4] nber.org, " Interest Rate" Accessed April 12, 2018 http://www.nber.org/papers/w234

[5] scb.se "consumer-price-index-cpi " Accessed April 12, 2018. http://www.scb.se/en/finding-statistics/statistics-by-subject-area/prices-and-consumption/consumer-price-index/consumer-price-index-cpi

[6] focus-economics.com, "/industrial-pro" Accessed April 12, 2018. https://www.focus-economics.com/economic-indicator/industrial-pro

[7] Lexicon.ft.com, "government-consumption" Accessed April 12, 2018. http://lexicon.ft.com/Term?term=government-consumption

[8] Focus-economics.com, "Private-consumption" Accessed April 12, 2018. https://www.focus-economics.com/economic-indicator/Private-consumption

[9] Investopedia.com , "Gross Domestic Production" Accessed April 12, 2018. https://www.investopedia.com/terms/g/gdp

[10] Dowjones.com , "Dow-jones " Accessed April 12, 2018. https://www.dowjones.com/

[11] Nasdaq.com , "NASDAQ" Accessed April 12, 2018. https://www.nasdaq.com/markets/indices/

[12] Quotes.wsj.com, "NIFTY50" Accessed April 12, 2018. https://quotes.wsj.com/index/IN/NIFTY50

[13] Quora.com, "Bank-Nifty" Accessed April 12, 2018. https://www.quora.com/What-does-Bank-Nifty-mean-in-the-Indian-stock-market

[14] stockmaster.in, "cnx-infrastructure-index." Accessed April 12, 2018. http://www.stockmaster.in/cnx-infrastructure-index.

[15] stockmaster.in, “cnx-100” Accessed April 12, 2018. http://www.stockmaster.in/cnx-100-index.html

[16] investopedia.com, “Hang-Seng” Accessed April 12, 2018. https://www.investopedia.com/terms/h/hangseng.asp

[17] investopedia.com, “DAX” Accessed April 12, 2018. https://www.investopedia.com/terms/d/dax.asp

[18] discovernikkei.org, “what-is-nikkei” Accessed April 12, 2018. http://www.discovernikkei.org/en/about/what-is-nikkei

[19] lexicon.ft.com, “FTSE-100” Accessed April 12, 2018. http://lexicon.ft.com/Term?term=FTSE-100

[20] Appel, Gerald. *Technical analysis: power tools for active investors*. FT Press, 2005.

[21] Wilder, J. Welles. *New concepts in technical trading systems*.Trend Research, 1978.

[22] Pring, Martin J. *Technical analysis explained*. McGraw-Hill Companies, 2002.

[23] StockCharts.com - ChartSchool , "Bollinger Bands." Accessed November 12, 2016 http://stockcharts.com/school/doku.php? id=chart_school:technical_indicators:bollinger_bands.

[24] StockCharts.com - ChartSchool , "Stochastic Oscillator." Accessed

November 12, 2016. http://stockcharts.com/school/doku.php?

[25] StockCharts.com - ChartSchool, "Commodity Channel Index (CCI)."Accessed October 5, 2012.http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:commodity_channel_index_cci.

[26] StockCharts.com - ChartSchool , "Chaikin Oscillator." Accessed October5, 2012.http://stockcharts.com/school/doku.php?

[27] StockCharts.com - ChartSchool , "Rate of Change (ROC)." AccessedOctober 5, 2012.http://stockcharts.com/help/doku.php?id=chart_school:technical_indicators:rate_of_change_roc_a.

[28] Hall, Mark A., and Lloyd A. Smith. "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper." *FLAIRS conference*.Vol. 1999. 1999.

[29] Yu, Lean, et al. "Evolving least squares support vector machines for stock market trend mining." *IEEE Transactions on evolutionary computation* 13.1 (2009): 87-102.

[30] Kim, Kyoung-jae. "Financial time series forecasting using support vector machines." *Neurocomputing* 55.1-2 (2003): 307-319.

[31] Tsai, C. F., and S. P. Wang. "Stock price forecasting by hybrid machine learning techniques."Proceedings of the International MultiConference of Engineers and Computer Scientists.Vol. 1.No. 755. 2009.

[32] Das, Shom Prasad, and SudarsanPadhy. "Support vector machines for prediction of futures prices in Indian stock market." International Journal of Computer Applications 41.3 (2012).

[33] Haider Khan, Zabir, TasnimSharminAlin, and AkterHussain."Price Prediction of Share Market Using Artificial Neural Network'ANN'."*International Journal of Computer Applications* 22.2 (2011): 42-47.

[34] Adebiyi, A. A., et al. "Stock price prediction using neural network with hybridized market indicators." *Journal of Emerging Trends in Computing and Information Sciences* 3.1 (2012): 1-9.

[35] Teixeira, Lamartine Almeida, and Adriano Lorena Inacio De Oliveira. "A method for automatic stock trading combining technical analysis and nearest neighbor classification."*Expert systems with applications* 37.10 (2010): 6885-6890.

[36] Huang, Wei, YoshiteruNakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine."*Computers & Operations Research* 32.10 (2005): 2513-2522.

[37] Tay, Francis EH, and Lijuan Cao. "Application of support vector machines in financial time series forecasting." *Omega* 29.4 (2001): 309-317.

[38] Thissen, U., et al. "Using support vector machines for time series prediction." *Chemometrics and intelligent laboratory systems* 69.1-2 (2003): 35-49.

[39] Linoff, Gordon S., and Michael JA Berry. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011.

[40] Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern classification*.Vol. 2. New York: Wiley, 1973.

[41] Shen, Shunrong, Haomiao Jiang, and Tongda Zhang. "Stock market forecasting using machine learning algorithms."*Department of Electrical Engineering, Stanford University, Stanford, CA* (2012): 1-5.

[42] Karazmodeh, M., S. Nasiri, and S. MajidHashemi. "Stock price forecasting using support vector machines and improved particle swarm optimization." *Journal of Automation and Control Engineering* 1.2 (2013): 173-176.

[43] Soni, Sneha, and ShailendraShrivastava. "Classification of Indian stock market data using machine learning algorithms."*International Journal on Computer Science and Engineering* 2.9 (2010): 2942-2946.

[44] Narayanan, B., and M. Govindarajan. "Prediction of Stock Market using Ensemble Model."*International Journal of Computer Applications* 128.1 (2015): 18-21.

[45] Dietterich, Thomas G. "Ensemble learning." *The handbook of brain theory and neural networks* 2 (2002): 110-125.

[46] Trybula, Walter J. "Data mining and knowledge discovery." *Annual review of information science and technology (ARIST)* 32 (1997): 197-229.

Appendix

Appendix A: Comparison of the models with Anova test

The Anova test was done on the four performance metrics for 10 fold cross validation and separate train and test. The following are the results from Anova test.

I. Mean Absolute Error

Table 9 Individual Prediction ability (10 fold validation)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	638.91	127.78	5.38	0.0320
Error	6	142.62	10.71		
Corrected Total	11	781.54			
R-Square	0.81				

Table 10 Redundant prediction (10 fold validation)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	592.60	118.52	11.06	0.0055
Error	6	64.28	23.77		
Corrected Total	11	656.89			
R-Square	0.902				

Table 11Individual Prediction ability (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	3813.18	762.63	4.50	0.0474
Error	6	1016.77	169.46		
Corrected Total	11	4829.96			
R-Square	0.789				

Table 12 Redundant prediction (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	3813.18	762.63	4.50	0.0474
Error	6	1016.77	169.46		
Corrected Total	11	4829.96			
R-Square	0.789				

Table 13 Mean separation for mean absolute error

	10 fold cross validation		Separate train and test set	
Machine learning Models	Individual Prediction MAE	Less redundant Features MAE	Individual Prediction MAE	Less redundant Features MAE
Ensemble	5.82[a]	6.89[a]	4.94[a]	4.94[a]
ANN	6.19[a]	6.20[a]	2.81[a]	2.81[a]
SVM	6.50[a]	9.84[a]	8.12[a]	8.12[a]
KNN	20.84[b]	21.32[b]	45.34[b]	45.34[b]
Mean	9.84	11.06	15.30	15.30
Minimum Significant Difference @ α=0.05	13.78	9.25	36.794	36.794

Means with separate letters denote significantly different models.

The analysis of variance indicated that the mean absolute error of machine learning models, *viz.*, ANN, Ensemble and SVM are not significantly different at α=0.05. Contrastingly, the mean absolute error of the K-NN prediction model was significantly higher as compared to the other models used in the study.

II. Root Mean Squared Error

Table 14 Individual Prediction ability (10 fold validation)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	1140.36	228.07	5.05	0.0369
Error	6	271.20	45.20		
Corrected Total	11	1411.56			
R-Square	0.807				

Table 15Individual Prediction ability (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	1009.66	201.93	10.47	0.0063
Error	6	115.67	19.27		
Corrected Total	11	1125.33			
R-Square	0.897				

Table 16 Individual Prediction ability (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	5827.48	1165.49	5.27	0.0335
Error	6	1327.87	221.31		
Corrected Total	11	7155.35			
R-Square	0.814				

Table 17 Redundant prediction (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	4844.22	968.84	14.93	0.0025
Error	6	389.23	64.87		
Corrected Total	11	5233.45			
R-Square	0.925626				

Table 18 Mean separation for root mean squared error

	10 fold cross validation		Separate Train and test	
Machine learning Models	Individual Prediction RMSE	Less redundant Features RMSE	Individual Prediction RMSE	Less redundant Features RMSE
Ensemble	9.3851[a]	11.8550[a]	7.4030[a]	17.7817[a]
ANN	10.8975[a]	11.2810[a]	6.0032[a]	13.5099[a]
SVM	10.6510[a]	15.1848[a]	12.6603[a]	16.7925[a]
KNN	29.1056[b]	30.9341[b]	58.3722[b]	62.0364[b]
Mean	15.0098	17.313725	21.109675	27.530125
Minimum Significant Difference @ α=0.05	19.003	12.411	42.048	22.765

Means with separate letters denote significantly different models.

The analysis of variance indicated that the root mean squared error of machine learning models, *viz.*, ANN, Ensemble and SVM are not significantly different at α=0.05. Contrastingly, the mean absolute error of the K-NN prediction model was significantly higher as compared to the other models used in the study.

III. Relative Absolute Error

Table 19 Individual Prediction ability (10 fold validation)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	1140.36	228.07	5.05	0.0369
Error	6	271.20	45.20		
Corrected Total	11	1411.56			
R-Square	0.807				

Table 20 Individual Prediction ability (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	37.11	7.42	3.67	0.0722
Error	6	12.11	2.01		
Corrected Total	11	49.231			
R-Square	0.753				

Table 21 Individual Prediction ability (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	140.91	28.18	9.41	0.0083
Error	6	17.96	2.99		
Corrected Total	11	158.87			
R-Square	0.886				

Table 22 Redundant prediction (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	233.20	46.64	3.69	0.0715
Error	6	75.80	12.63		
Corrected Total	11	309.01			
R-Square	0.754				

Table 23 Mean separation for relative absolute error

	10 fold cross validation		Separate Train and test	
Machine learning Models	Individual Prediction RAE	Less redundant Features RAE	Individual Prediction RAE	Less redundant Features RAE
Ensemble	1.2281[a]	1.4631[a]	0.9261[a]	2.0087[a]
ANN	1.1570[a]	1.2082[a]	0.4926[a]	1.0666a
SVM	1.1083[a]	1.7546[a]	1.6917[a]	4.2129[a]
KNN	7.1015[b]	5.4213[b]	8.6460[b]	10.4993[b]
Mean	2.648725	2.4618	2.9391	4.446875
Minimum Significant Difference @ α=0.05	5.5474	4.0171	4.8908	10.047

Means with separate letters denote significantly different models.

The analysis of variance indicated that the relative absolute error of machine learning models, *viz.*, ANN, Ensemble and SVM are not significantly different at α=0.05. Contrastingly, the mean absolute error of the K-NN prediction model was significantly higher as compared to the other models used in the study.

IV. Root Relative Squared Error

Table 24 Individual Prediction ability (10 fold validation)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	88.65	17.73	4.92	0.0389
Error	6	21.61	3.60		
Corrected Total	11	110.26			
R-Square	0.804				

Table 25 Individual Prediction ability (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	41.16	8.23	3.57	0.0764
Error	6	13.83	2.30		
Corrected Total	11	54.99			
R-Square	0.748				

Table 26 Individual Prediction ability (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	169.86	33.97	25.65	0.0006
Error	6	7.94	1.32		
Corrected Total	11	177.80			
R-Square	0.95				

Table 27 Redundant prediction (Separate train and test)

Source	DF	Sum of squares	Mean square	F value	Pr>F
Model	5	260.00	52.00	5.69	0.0281
Error	6	54.85	9.14		
Corrected Total	11	314.85			
R-Square	0.82				

Table 28 Mean separation for root relative squared error

	10 fold cross validation		Separate Train and test	
Machine learning Models	Individual Prediction RRSE	Less redundant Features RRSE	Individual Prediction RRSE	Less redundant Features RRSE
Ensemble	1.6656[a]	2.1911[a]	1.1769[a]	3.0654a
ANN	1.8890[a]	2.1050[a]	0.9434[a]	1.0666a
SVM	1.8888[a]	2.7224a	2.3074[a]	5.1857a
KNN	7.9802[b]	6.4781[b]	9.9845[b]	11.5133[b]
Mean	3.3559	3.37415	3.60305	5.20775
Minimum Significant Difference @ α=0.05	5.3641	4.292	3.253	8.5465

Means with separate letters denote significantly different models.

The analysis of variance indicated that the root relative squared error of machine learning models, *viz.*, ANN, Ensemble and SVM are not significantly different at α=0.05. Contrastingly, the mean absolute error of the K-NN prediction model was significantly higher as compared to the other models used in the study.