Develop Model on Market Manipulation for Ethiopian Commodity Exchange Using Machine Learning Manipulation

A Thesis Presented

by

Biniam Gebremedhin

to

The Faculty of Informatics

of

St. Mary's University

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

in

Computer Science

June 2024

ACCEPTANCE

Develop Model on Market Manipulation for Ethiopian Commodity Exchange Using Machine Learning Manipulation

By

Biniam G/medhin

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

_____

Internal Examiner

Dr. Alembante Mulu

_____

External Examiner

_____

Dean, Faculty of Informatics

Dr. Alembante Mulu

June 2024

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Biniam G/medhin

Signature

_____

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor.

Alembante Mulu

Signature

_____

Addis Ababa

Ethiopia

June 2024

Acknowledgment

This study on constructing a model for market manipulation on the Ethiopian Commodity Exchange (ECX) using machine learning would not have been feasible without the assistance and participation of several individuals and organizations.

Primarily, I would want to convey my heartfelt appreciation to Dr. Alembante, my academic adviser, whose advice, insight, and support were invaluable throughout the study process. Your experience and steadfast support have been helpful.

I would also want to thank the instructors and staff at St. Mary's University for providing the resources and a favorable environment for study. Special thanks to the Department of Computer Science for their ongoing assistance.

I am very thankful to the Ethiopian Commodity Exchange (ECX) for providing access to the relevant data and cooperating throughout this investigation. The views and input from ECX experts were critical in establishing the direction and emphasis of this study.

Thank you to my colleagues and fellow researchers for providing helpful input and sharing their expertise and experience. Your contributions have greatly enriched this work.

Finally, I would want to thank my family and friends for their support and understanding. Your compassion and support have helped me stay strong and motivated during this journey.

Contents

List of Tables

List of Figures

List of Abbreviations

AI	**Artificial Intelligence**
ANN	**Artificial Neural Networks**
CSRC	**China Securities Regulation Commission**
CNN	**Convolutional Neural Network**
DTC	**Decision Tree Classifier**
DL	**Deep Learning**
e-trade	**Electronic Trade**
ECX	**Ethiopia Commodity Exchange**
GAN	**Generative Adversarial Networks**
KNN	**K-nearest Neighbor**
LR	**Linear Regression**
LR	**Logistic Regression**
LSTM	**Long Short Term Memory**
ML	**Machine Learning**
MoT	**Ministry of Trade**
RBF	**Radial Basis Functions**
RF	**Random Forest**
RFDT	**Random Forest Decision Tree**
RNN	**Recurrent Neural Network**
SNN	**Simulated Neural Networks**
GBM	**Gradient Boost Machine**

Abstract

Market manipulation poses a significant threat to the integrity and efficiency of financial markets and commodity markets, particularly in emerging markets such as the Ethiopian Commodity Exchange (ECX). This thesis aims to develop a robust machine-learning model to detect and mitigate market manipulation within the ECX. By leveraging historical transaction data and employing advanced machine learning algorithms, the study seeks to identify anomalous trading patterns indicative of manipulative activities.

The research begins with a comprehensive review of the existing literature on market manipulation detection and machine learning techniques. Subsequently, a detailed analysis of the ECX's trading data is conducted to understand the unique characteristics and potential vulnerabilities of this market. Data preprocessing techniques are employed to cleanse and prepare the data for model training.

Various machine learning models, including supervised and unsupervised learning algorithms, are evaluated for their efficacy in detecting market manipulation. The models are trained on labeled datasets containing instances of known manipulative activities and normal trading behavior. Performance metrics such as accuracy, precision, recall, and F1-score are used to assess the effectiveness of each model.

The results demonstrate that certain machine learning models, particularly ensemble methods and neural networks, show high potential in accurately detecting market manipulation within the ECX. The best-performing model is integrated into a real-time monitoring system, providing timely alerts to market regulators and stakeholders.

This study contributes to the body of knowledge by offering a novel approach to market manipulation detection in commodity exchanges, with a specific focus on the Ethiopian context. The developed model not only enhances market surveillance capabilities but also promotes market integrity, investor confidence, and overall market stability.

Future work will explore the scalability of the model to other emerging markets and the incorporation of additional data sources, such as social media sentiment and economic indicators, to further enhance the model's predictive power.

CHAPTER ONE

1. Introduction

1.1. Background

Market surveillance is the way of detecting and finding out abnormal trade in the commodity market or stock market. The prevention and examination of manipulative, abusive, or unlawful trading behaviors in the securities markets is known as market surveillance. When buyers and sellers are confidence in the fairness and accuracy of transactions, they are more eager to engage in orderly markets, which are maintained by market monitoring. A market that is not monitored might becoming chaotic, which would deter investment and impede economic expansion. Both the public and private sectors are able to offer market surveillance [1].

Market surveillance, as defined by Kento, refers to illicit trading that takes place between a buyer and a seller. The term "market manipulation" covers a broad spectrum of trade activities that inflate prices, provide market manipulators with an advantage over other players, and produce information asymmetries [2]. So any behavior that distort price in both commodity and stock market will be referred to us a surveillance issue.

Market monitoring systems serve an important role in avoiding misbehavior and violations of trade norms, which extends beyond the direct requirement of regulatory organizations [3].

Establishing a market surveillance department to monitor all trades using a set of standard operating procedures that the company deems manipulative is mandatory for any corporation dealing in stocks or commodities.

Market integrity is the capacity of players to deal in a fair and informed market, where prices reflect information [4]. Therefore, it appears that fairness in this case may be limited to making sure the markets are impartial and equitable, while integrity can be limited to making sure the markets are "unimpaired," "uncorrupted," and "sound." [5].

The primary player in market integrity is market surveillance. With its rule definition, the market surveillance department portrays any potential trade with integrity issues. The term "market

manipulation" covers a broad spectrum of trade activities that inflate prices, provide market manipulators with an advantage over other players, and produce information asymmetries [2]. Exchanges and securities commissioners conduct market surveillance to identify market manipulation by market players.

Additionally, a number of methods of manipulating the market are delineated, including pumping and dumping, flash crushing, insider trading, spoofing, painting the tape, ramping, and so forth. Because of this, it might take a long time to implement all of these surveillance issues on each trade, which would put a lot of pressure on the surveillance officer and result in a lot of market manipulation.

The Ethiopia Commodity Exchange is a relatively new addition to Ethiopia and the first business of its sort in Africa. The ECX is a special collaboration between market participants and the exchange's members, primarily supported by the Ethiopian government. Ethiopia's future depends on ECX since it provides the market's integrity, security, and efficiency. ECX presents unprecedented development potential in the commodities industry and related businesses like as transportation and logistics, banking and financial services, and others [6].

The Ethiopia Commodity Exchange is an Ethiopian spot exchange based in Addis Ababa. Through the open outcry trading mechanism and electronic trade (e-trade), ECX members or their authorized agents trade over 200 different spot contracts [6].

The exchange also includes a market surveillance department that ensures the integrity and liquidity of spot trading. Using a pre-established market manipulation rule that the exchange has already established, this department closely monitors each transaction that takes place on the spot market and responds promptly to any completed trades.

The exchange also keeps an eye on the integrity of the market using its rule-based trade surveillance mechanism. This division deals with any trade manipulation that the exchange defines as manipulation based on the price and quantity provided by the traders. If any manipulation occurs during a deal, the surveillance officer takes immediate action by thoroughly researching the situation and initiating legal proceedings.

During a trade, many data will be generated; it will take a lot of effort and time to distinguish between a legitimate trade and an abusive transaction. Additionally, this anomaly report will

influence trade integrity and liquidity by yielding more false positive results than real positive results. Additionally, by using the defined rule to identify a false positive, the trader and surveillance officer waste time.

This research suggests a trade manipulation model that overcomes all of the aforementioned difficulties. There are numerous methods for creating a fully scaled and end-to-end commodity trading and stock market monitoring system. We propose a model that detects anomalous trades and sends a notification to the surveillance officer.

With a wide range of applications, machine learning is one of the fields of computer science with the most rapid development. It describes the automatic identification of significant patterns in data [7].

A machine learning approach will be implemented to predict and classify a trade manipulation. Artificial Intelligence (AI) and Machine Learning technologies provide a route toward less expensive and more accurate trade surveillance. This research paper provides an outline for the transition to AI-based technologies, explains the arising opportunities, and sheds light on associated challenges and approaches to overcome them by highlighting the various use cases and benefits of integrating Machine Learning techniques into trade surveillance. As technology evolves, it is critical to stay up by building advanced monitoring systems capable of detecting and alerting to market manipulation. Such a monitoring system might benefit from a machine learning technique.

Machine learning and AI-based technologies give new opportunities for more effective and efficient trade surveillance tactics. These solutions benefit both regulators (Ethiopian Commodity Exchange Authority) and organizations (Ethiopia Commodity Exchange) that use surveillance. It is critical to understand the underlying processes, vendor solutions, and, eventually, the result goals in such a change to cutting-edge technology through clearly defined phases. Effective training enables a clear and transparent information flow between compliance officers and authorities.

1.2. Motivation

Market manipulation damages the exchange's fairness and credibility. Creating a machine-learning model to detect and prohibit such manipulations contributes to a fair playing field for all participants. Furthermore, ensuring that the market functions fairly builds trust among investors, traders, and other stakeholders, which is critical for the exchange's development and stability. In times of economic stability, market manipulation can result in artificial price variations that hurt both producers and consumers. By detecting and preventing such activities, the model contributes to the maintenance of stable and equitable commodity prices, which is critical for economic stability. In addition, regulatory organizations such as the Ministry of Trade (MOT) and others are mandated to detect and prevent market manipulation in order to guarantee equitable trade practices. Creating a machine learning model helps the ECX adhere to legal norms and these regulatory obligations. Sophisticated machine learning models offer sophisticated capabilities for keeping an eye on trade activity, making regulatory supervision more effective and efficient.

In case of technology as we mentioned above traditional techniques of identifying market manipulation may have limited scope and accuracy. Machine learning models are capable of analyzing large volumes of data, identifying complicated patterns, and improving detection accuracy.

1.3. Statement of the Problem

Many electronics trades were completed in a single day on the Ethiopian Commodity Exchange. Distinguishing between normal and manipulative trades is time-consuming and error-prone. As a result, false positive trade is treated as regular trade. One of the main goals of exchanges is to maintain the integrity and efficiency of the market, which is essential to its future competitiveness.

The exchange employs its own logic (rules) to identify manipulative trades from non-manipulative ones carried out by the surveillance department. This technique requires a large amount of time or takes a day to detect the abusive trade.

The Ethiopia Commodity Exchange must use some form of automated market manipulation mechanism that employs machine learning or other AI-based technologies.

If the exchange fails to address the issue in a systematic manner, the market's integrity would deteriorate, perhaps leading to an economic catastrophe in the country's financial markets.

Therefore, this research attempts to fill this gap and contributes to the Ethiopian Commodity Exchange surveillance department a model using machine learning on market manipulation.

1.1. Research Questions

Those are the research questions that could guide the study on developing a model for detecting market manipulation on the Ethiopian Commodity Exchange using machine learning:

1. How can trade data from the Ethiopian Commodity Exchange be used to build machine-learning models for early market manipulation detection and classification?
2. Which machine learning techniques are suitable for detecting anomalous transaction patterns on the Ethiopian Commodity Exchange? How can these methodologies indicate market manipulation?
3. To what extent can feature engineering increase the effectiveness and interpretability of machine-learning models for detecting and categorizing market manipulation on the Ethiopian Commodity Exchange?

These research questions cover different aspects of developing a model for detecting market manipulation using machine learning. Each question aims to explore a specific facet that is crucial for the successful development and implementation of such a model in the context of the Ethiopian Commodity Exchange.

1.2. Objectives

1.2.1. General Objective

The model will be evaluated using strong performance indicators such as accuracy, precision, recall, and the F1 score. These metrics will assess how successfully the model recognizes potential cases of market manipulation. The results aim to increase openness and confidence in commodities trading in Ethiopia's rural economy by providing helpful information to ECX authorities and stakeholders in order to improve supervision and regulatory measures.

This study contributes to the larger goal of protecting market integrity and fostering fair trading practices in Ethiopian commodity markets by providing an effective model for identifying ECX market manipulation.

1.2.2. Specific Objective

In relation to the above general objective of the study, the researcher had identified the following specific objectives:

- To recognize important characteristics and signs of market manipulation
- To collect and preprocess relevant data
- To develop and implement machine learning models on market manipulation on Ethiopian Commodity Exchange
- To review state of the art literature review on market manipulation on Ethiopian Commodity Exchange
- To prepare the required dataset on trade manipulation
- To study the factor that contribute on market price or quantity manipulation
- To identify suitable features or parameters on market manipulation
- To evaluate the performance of the machine learning models on Ethiopian Commodity Exchange
- To develop a comprehensive detection framework

1.3. Significance of the study

This research paper, develops a machine learning model for detecting market manipulation on the ECX, has significant value in multiple dimensions, including market integrity, participant protection, economic benefits, technological advancement, educational contributions, policy support, social impact, and long-term sustainability. These benefits add up to a more strong, transparent, and efficient commodities market in Ethiopia.

1.4. Scope/limitations

As previously said, market manipulation differs among organizations dependent on their country's laws and regulations. This research only includes trade manipulation as defined by the Ethiopia Commodity Exchange and the Ministry of Trade. In this research work, we utilize a machine learning strategy to include all ECX and MOT trade manipulation.

1.5. Organization of the rest of the thesis

Creating a paper to construct a model for market manipulation on the Ethiopian Commodity Exchange (ECX) using machine learning entails arranging it in a way that clearly conveys the research process, findings, and consequences. The second part of this paper is all about the literature review, the third chapter is about methodology, the fourth chapter is about design, execution, experimental findings, and discussions, and the fifth chapter is about the conclusion.

CHAPTER TWO

2. Literature

2.1. Introduction

Market surveillance is a technique for detecting fraudulent financial theft on any financial institution that executes a trade between two parties, the buyer and seller.

Market surveillance is essential for ensuring the integrity and liquidity of the country's financial flow in any commodities market or financial institution. As a result, any financial institution's market surveillance system must be powerful and technologically advanced in order to avoid being compromised by any unauthorized trader.

If we use a standard rule-based technique for market manipulation, today's market surveillance for commodities exchange becomes more difficult. Market manipulation has gotten more sophisticated as the demand for commodities has increased.

This study presents an in-depth examination of cutting-edge machine learning approaches utilized in financial market manipulation.

We discuss the challenges and advancements in this field of research, especially in relation to other application domains. In this research study, we emphasize the design of a machine learning-based surveillance system for a commodities trading market and investigate how the nature of the input data affects the effectiveness of the methods for spotting abused market behaviors. Overall, our findings show that the algorithms we employ anticipate day-ahead future prices accurately and successfully, proving their ability to spot abused price fluctuations.

Exchanges are for-profit companies that operate as marketplaces for the trading of securities and other market contracts. The operations of an exchange's surveillance departments are performed at a cost to the exchange; in other words, these departments operate as "cost centers." [8]. Ethiopia Commodity Exchange manages trade manipulation issue using its surveillance department.

What is trade manipulation? The term "market manipulation" covers a broad spectrum of trade activities that inflate prices, provide market manipulators with an advantage over other players, and produce information asymmetries [8].

When monopolistic power is used in the futures market or the cash market for the underlying commodity close to a futures contract's expiration date, it is referred to as manipulation [9].

The primary argument is that manipulation affects pricing, resulting in inefficient resource allocation, which the right objective of government policy is. Manipulation is nearly usually done for financial advantage, and hence it is a type of rent seeking: resources are used to obtain a private profit [10].

The Ethiopian Commodity Exchange enumerates some forms of manipulation that may affect the commodity market's fairness and liquidity.

Pre-Arranged Trade

Trade that occurs at predetermined, mutually agreed-upon prices before to execution is referred to as pre-arranged trade.

Cancellation Frequency

It refers to the frequency with which a potential match deal is canceled; this behavior is portrayed as trade manipulation in the Ethiopia commodity exchange.

Front Running

Based on insider information that their company is likely to propose a purchase or sell to customers, who would almost surely influence the price of a commodity, a representative (broker) may also front-run.

Match Trade

When scammers use matched orders to manipulate the market, they initiate transactions to buy or sell commodities knowing that will or have already been a matching order made on the other side.

Trade between affiliated/sister companies, wash, or fictions trade

When a trade occurs between related or sibling firms the system ought to raise an alert by flashing a flag based on membership data and the data we pass in.

Trade Spoofing

Spoofing Trading, also known as bluffing or trade spoofing, is a disruptive trading technique in which a trader submits a huge bid or ask order to the market with the goal of canceling the order before execution in order to influence the market. The enormous order dupes other traders into believing that there is a high demand or supply in the market and trading accordingly.

Flash Crash

A flash crash is a type of financial event when stock orders or sales are quickly withdrawn; causing a sharp decline in price that is usually followed by a quick rebound in a few minutes or hours, usually on the same day. Electronic securities markets typically experience something similar to this.

These are a few examples of the key manipulation issues that arise in daily trading activities on the Ethiopia Commodity Exchange.

2.2. Approaches

Artificial intelligence is come up with deep learning and machine learning this two are refer to as a sub-field of artificial intelligence. Because deep learning and machine learning are often used interchangeably, it is important to understand the differences between the two. As previously stated, both deep learning and machine learning are subfields of artificial intelligence, and deep learning is essentially a sub-field of machine learning [11].

The main difference between deep learning and machine learning is that non-deep machine learning requires human interaction during subsequent data extraction deep learning uses automated methods for data extraction.

Deep learning and machine learning differ in how their algorithms learn. Deep learning automates most of the feature extraction process, minimizing the need for manual human involvement and allowing the usage of bigger data sets. Classical or "non-deep" machine learning relies more on human assistance to learn. Human specialists develop the hierarchy of features to grasp the distinctions between data inputs, which often require more structured data to learn [11].

From the above statement, we clearly put the main difference between artificial intelligence, machine learning and deep learning. Now let us look which approaches or techniques used for the solving of our problem domain.

MACHINE LEARNING

Machine learning is a subset of artificial intelligence that is used to forecast, categorize, image - processing, language processing, and so on. It works by providing a data, training from it, and providing a correct prediction value. Sometimes after viewing the data, we cannot interpret the extract information from the data.

Having so much data available, the demand for machine learning is growing. Machine learning is utilized in a variety of industries to retrieve valuable data. Machine learning is intended to learn from data. Many studies have been carried out to establish how to train computers to learn on their own without being explicitly programmed. Many mathematicians and programmers utilize various methods to handle this problem with enormous data sets [7].

Machine learning addresses data difficulties using a number of ways. Data scientists want to underline that there is no single algorithm that performs effectively in all situations [12]. The type of method employed is determined by the type of problem you are attempting to solve, the number of variables involved, the appropriate model to apply, and other criteria. Here is a quick introduction of some of the most widely used machine learning (ML) algorithms.

How machine learning work?

Machine learning work lays on three pillars decision process, error process and modeling optimization. Decision Process: a process on which data prediction or classification based on labeled data or unlabeled data.

Error Process: An error function evaluates the prediction of the model and gives us the accuracy of the model that we defied.

Modeling Optimization: If the model fits the data points in the training set better, the weights are modified to lessen the difference between the known example and the model prediction. The algorithm will repeat this "evaluate and optimize" procedure, updating weights autonomously until an accuracy criterion is reached.

Machine Learning Methods

Machine learning models fall into three primary categories i.e. supervised machine learning, unsupervised machine learning and Semi-supervised learning.

Supervised Machine Learning

Supervised learning, commonly referred to as supervised machine learning, is a subfield of machine learning and artificial intelligence. It is defined by the use of labeled datasets to train algorithms for properly classifying data or predicting outcomes [13].

A function that converts an input to an output is taught through supervised learning with sample input-output pairs. It employs tagged training data consisting of a set of training samples to infer a function [14]. Supervised machine learning refers to algorithms that require outside assistance. The input dataset serves as the foundation for both the training and testing datasets. The output variable in the train dataset must be predicted or classified.

All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification. The workflow of supervised machine learning algorithms is given in figure below.

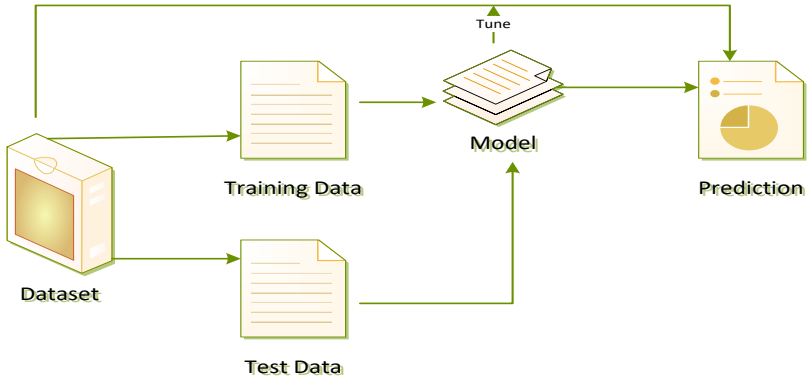

FIGURE 1 SUPERVISED LEARNING

Unsupervised Learning

Unsupervised learning is distinguished from supervised learning by the absence of right responses and instruction. The algorithms are permitted to discover and exhibit the remarkable structure in the data on their own [15]. Unsupervised learning methods extract a few properties from the data.

When new data is provided, it recognizes its class based on previously learnt properties. It is mostly used for feature reduction and grouping.

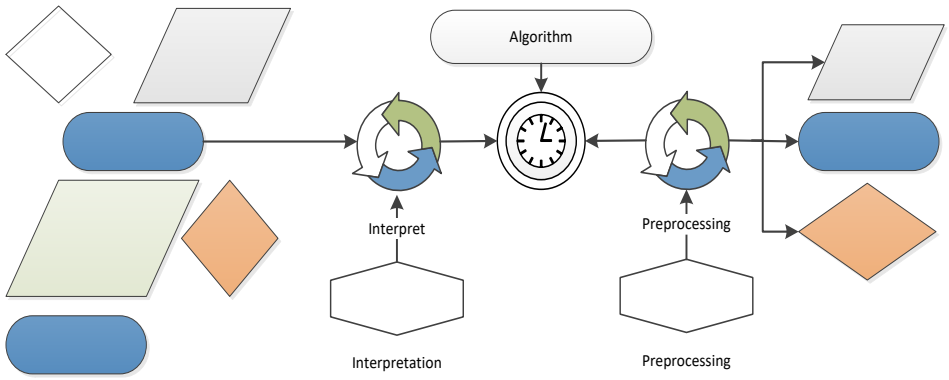

FIGURE 2 UNSUPERVISED LEARNING

The primary distinction between supervised and unsupervised learning is the type of input data used. Unlike unsupervised machine learning algorithms, supervised learning uses labeled training data to verify whether pattern recognition in a dataset is correct.

The objectives of supervised learning models are also specified, which means that the model's output is known before the algorithms are executed. In other words, the input is mapped to the output using the training data [16].

Common Supervised Machine Learning Algorithms

There are number of supervised machine learning algorithms are commonly used. These include:

Naive bayes:

The Naive Bayes classifier is a probabilistic algorithm that relies on the assumption of strong (naïve) independence across features, and is based on Bayes' theorem [17]. This means that the presence of one trait has no bearing on the presence of another on the likelihood of a certain occurrence, and that each predictor has the same influence on that result. Multinomial Nave Bayes, Bernoulli Nave Bayes, and Gaussian Nave Bayes are the three varieties of Nave Bayes classifiers. This approach is most typically applied in text classification, spam detection, and recommendation systems [18]. The Naïve Bayes classifier is a supervised machine-learning method commonly used for text categorization applications. It is also a member of a family of generative learning

algorithms, which means it, attempts to model the distribution of inputs for a certain class or category. Unlike discriminative classifiers such as logistic regression, it does not learn which characteristics are most essential to distinguish across classes [17].

Consider a supervised learning problem in which we wish to approximate an unknown target function $f: X \to Y$ or equivalently $P(X|Y)$. To begin, we will assume Y is a boolean-valued random variable, and X is a vector containing n boolean attributes. In other words, $X = (X1, X2, X3, \dots\dots\dots\dots Xn)$ where, Xi is the boolean random variable denoting the ith attribute of X.

Applying Bayes rule, we see that $P(Y = yi|X)$ can be represented as

$$P(Y = y_i|X = x_k) = \frac{P(x = X_k|Y = y_i)P(Y = y_i)}{\Sigma_{jP}\big(X = x_k|Y = y_j\big)P\big(Y = y_j\big)}$$

Linear regression:

Linear regression is a typical method for predicting future outcomes by detecting the relationship between a dependent variable and one or more independent variables. Simple linear regression is employed when there is only one independent variable and one dependent variable. Multiple linear regressions are employed as the number of independent variables increases. Its goal is to illustrate a line of best fit for each type of linear regression using the least squares method [20]. However, unlike other regression models, when shown on a graph, this line is straight.

x:inputvariable(alsocalledindependent,predictor,explanatoryvariable)
y:outputvariable(alsocalleddependent,responsevariable) A scatter plot shows (xi,yi) for i=1... n as dots. Model assumption:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

In the example above, y is the dependent variable, and x1, x2, and so on, are the explanatory variables. The coefficients (β_0 , β_1, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. β_0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

Logistic regression:

Multiple predictor adjustments are possible using logistic regression, which also allows for the use of continuous or categorical predictors. Because of this, logistic regression is particularly helpful when analyzing observational data when adjustments are required to lessen the possibility of bias arising from variations in the groups under comparison [21].

As opposed to linear regression, which is used when the dependent variable is continuous, logistic regression is used when the dependent variable is categorical, which means it includes binary outputs such as "true" or "false" or "yes" or "no." While both regression approaches aim to uncover correlations between data inputs, logistic regression is generally applied to solve binary classification challenges like spam detection.

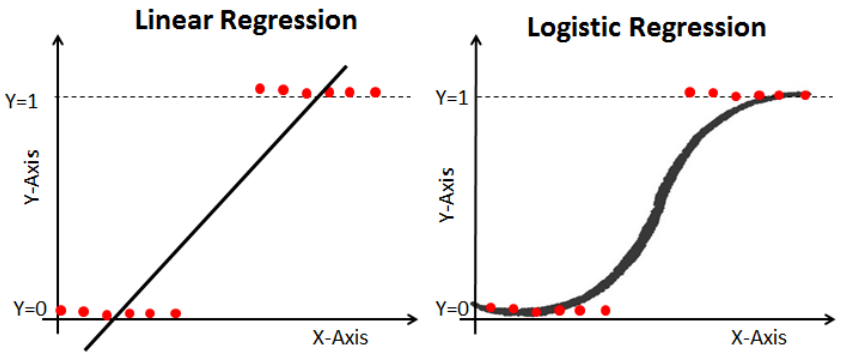

FIGURE 3 LOGISTIC REGRESSION

Gradient boost machine

In machine learning, boosting is a robust ensemble strategy. Boosting builds a single, more accurate strong learner by combining the predictions of several weak learners, in contrast to standard models that learn from the data separately. Gradient boosting has grown so prevalent in machine learning that it is being used in a wide range of sectors, from identifying asteroids to forecasting customer attrition [22].

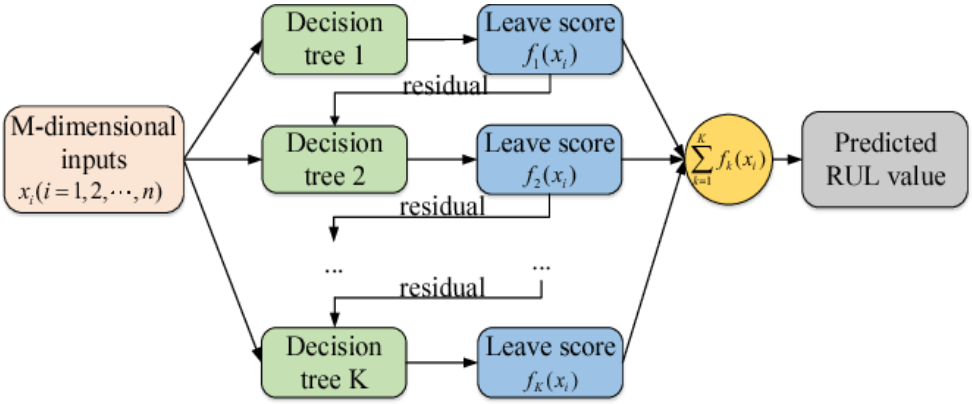

FIGURE 4 GRADIENT BOOSTING EXAMPLE

Random forest:

Random forest is another flexible supervised machine learning approach that may be applied to classification and regression. The word "forest" refers to a collection of uncorrelated decision trees that are then joined to reduce variance and create better data predictions.

Based on the concept of randomization, the Random Forest ensemble classifier creates a collection of independent, non-identical decision trees. $\{h(x,\theta_k), k = 1, \ldots, L\}$ Is a type of mutual independent random vector parameter, and x is the input data. This is the definition of random forest. Random vectors are used as parameters in each decision tree, and samples' features and the subset of the sample data set that serves as the training set are chosen at random [22].

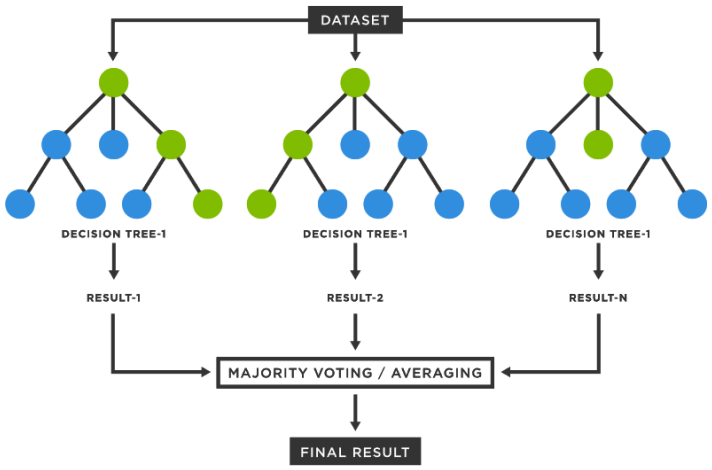

FIGURE 5 RANDOM FOREST

Decision Tree:

These models predict the value of a target variable by learning simple decision rules inferred from the data features. This approach for supervised learning is applied to classification issues. It performs effectively when categorizing dependent variables that are continuous or categorical. The feature is divided into two or more homogenous sets by this method according to the most important characteristics or independent variables.

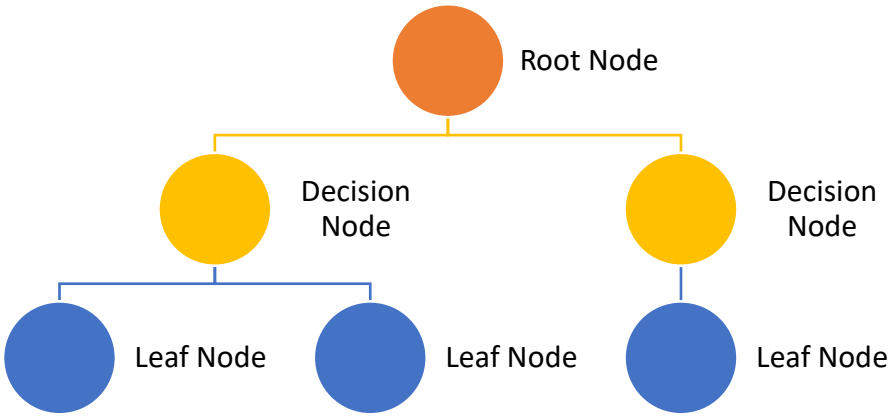

FIGURE 6 DECISION TREE

SVM (Support Vector Machine) Algorithm:

SVM is mostly used for classification but is also useful for regression in high-dimensional domains. Plotting raw data as dots in an n-dimensional space is how the SVM algorithm does classification. Classifying the data is made simple by assigning a coordinate to each feature's value. The data may be divided into groups and shown on a graph using lines known as classifiers [23].

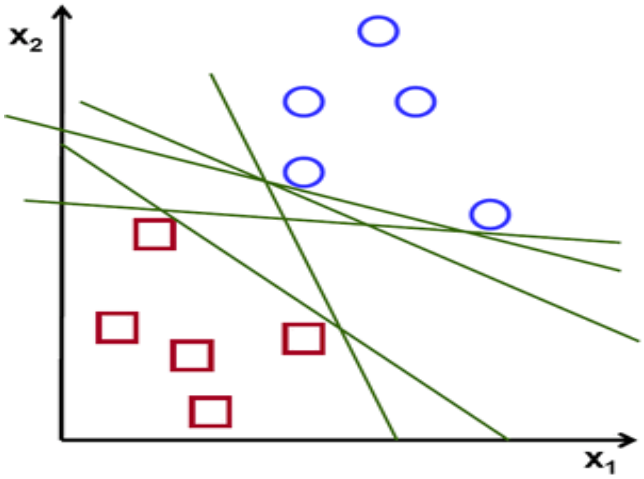

FIGURE 7 : SUPPORT VECTOR MACHINE SVM

Common unsupervised machine learning algorithms

There are number of unsupervised machine learning algorithms are commonly used. These include:

K-nearest neighbor:

The KNN algorithm, also known as K-nearest neighbor, is a non-parametric approach for categorizing data points based on their similarity to other accessible data. This approach thinks that data points with similar features can be found nearby. As a result, it attempts to calculate the distance between data points, typically using Euclidean distance, before assigning a category based on the most frequently occurring category or average. Its ease of use and low calculation time make it popular among data scientists but as the test dataset grows, so does the processing time, making it less suited for classification assignments. KNN is widely used in recommendation engines and image recognition.

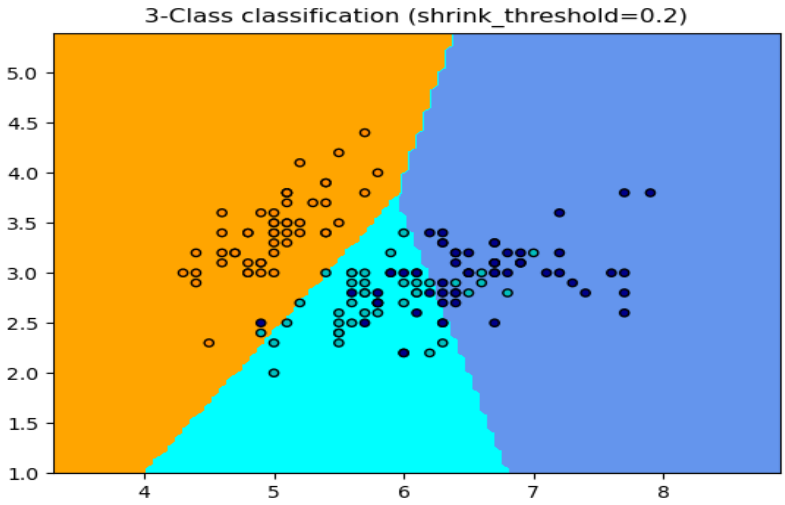

FIGURE 8 K-NEAREST NEIGHBORS KNN

K-mean

The K-Means approach clusters data by dividing samples into n groups with comparable variance while reducing inertia, also known as within-cluster sum-of-squares. This method requires that the number of clusters be specified. It scales efficiently to large numbers of samples and has been used in a variety of areas.

The k-means method separates a collection of N samples X into K disjoint clusters, each characterized by the mean μj of the samples in the cluster. The means, also known as cluster centroids, may not always correspond to points from X, although sharing the same space.

The K-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion:

$$\sum_{i=0}^{n} \min_{\mu j \epsilon C}(||xi - \mu j||)^2$$

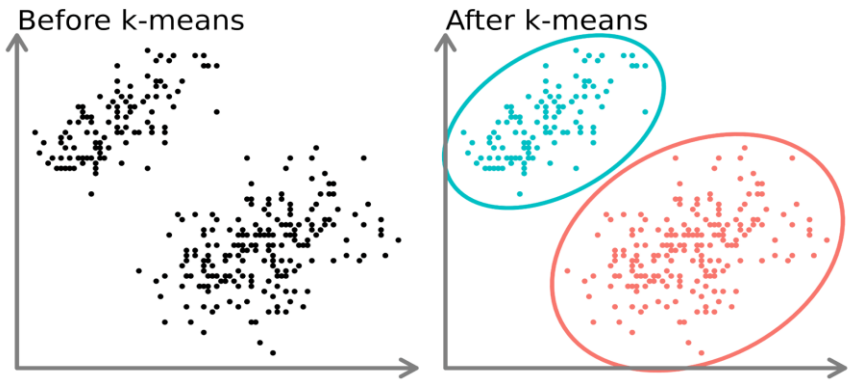

FIGURE 9 K-MEANS

DEEP LEARNING

Deep learning is a branch of machine learning that tries to learn high-level abstractions from data using hierarchical structures. It is a new method and has been widely employed in classic artificial intelligence fields, such as semantic parsing, transfer learning, natural language processing, computer vision, and many more [18].

Deep learning is a subtype of machine learning in which the neural network has three or more layers. These neural networks seek to replicate the function of the human brain, albeit far from its capabilities, allowing it to "learn" from enormous volumes of data. While a neural network with a single layer may still produce approximate predictions, more hidden layers can assist to enhance and refine the accuracy [19].

How deep learning work

Deep learning algorithms employ neural networks meant to mimic the human brain. Millions of connected neurons, for example, work together in the human brain to learn and process information. In a similar line, artificial neural networks, also known as deep learning neural networks, are made up of many layers of synthetic neurons that function together within a computer.

Artificial neurons are software components known as nodes that process data using mathematical algorithms. These nodes are employed by artificial neural networks and deep learning algorithms to solve complex problems.

Components of a deep learning network

In its most basic form, a deep learning contains three layers: input layer, hidden layer, and output layer.

Input Layer

Multiple nodes provide data into an artificial neural network. The system's input layer is composed of these nodes.

Hidden Layer

The neural network's input layer processes data before sending it to the next level. These hidden layers evaluate data at different levels and change their behavior in reaction to new data. Deep learning networks can investigate a problem from several viewpoints thanks to their hundreds of hidden layers.

Output Layer

The output layer consists of the nodes that output data. Deep learning models' output layer contains just two nodes that respond with "yes" or "no". In contrast, more nodes in those generate a wider range of responses.

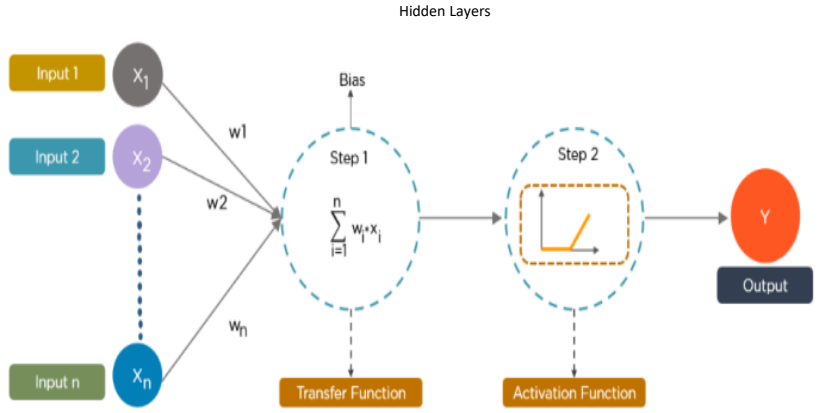

FIGURE 10 DEEP LEARNING

Common deep learning algorithms

Neural networks:

Neural networks, which are clearly built for deep learning algorithms, examine training data by mimicking the human brain's interconnections via node layers. Each node has inputs, weights, a bias (or threshold), and an output. If the output value hits a specific threshold, the node "fires," or activates, and sends data to the network's next layer. Neural networks learn this mapping function through supervised learning, modifying dependent on the loss function using gradient descent. When the cost function is near to zero, we may be certain that the model will generate the correct answer. Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning that form the basis of deep learning techniques. Their name and structure are inspired by the human brain, replicating the way organic neurons communicate with one another [20].

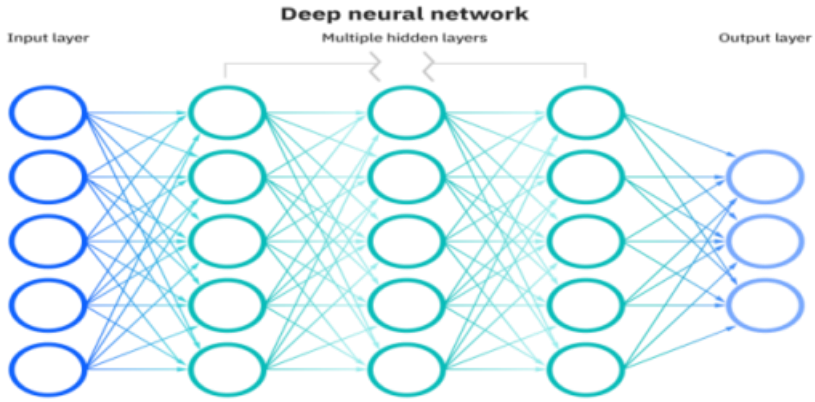

FIGURE 11DEEP NEURAL NETWORK

Consider every node to be a separate linear regression model, with input data, weights, a bias (or threshold), and an output. The calculation might resemble this:

$$\sum WiXi + bias = W1x1 + W2x2 + W3x3 + bias$$

$$Output = f(x) = 1\ id \sum X1x1 + b \geq 0\ if \sum X1x1 + b < 0$$

Deep learning models make use of many algorithms. Even while no network is perfect, various algorithms perform better for specific tasks than others. It is beneficial to have a solid understanding of each main algorithm in order to make the best judgments.

Convolutional Neural Networks (CNNs)

Convolutional matrices are frequently used in image processing to convolute through images and produce desired outputs.

Long Short Term Memory Networks (LSTMs)

An LSTM is a recurrent neural network (RNN) capable of learning and recalling long-term dependencies. The default choice is to save prior information for an extended period. LSTMs learn and retain information over time. Their ability to recall past inputs makes them effective in time series prediction. An LSTM consists of four interacting layers structured in a chain-like arrangement, allowing for unique communication. LSTMs are widely utilized in voice recognition, music creation, pharmacological research, and time-series prediction applications.

Recurrent Neural Networks (RNNs)

Given that of the directed cycle connections, LSTM outputs may be fed into the current phase of an RNN. Because of its internal memory, the LSTM's output serves as an input for the present phase while also recalling prior inputs. RNNs are extensively utilized in machine translation, natural language processing, handwriting recognition, time series analysis, and picture captioning.

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are deep learning algorithms that create new data by simulating the training set. A GAN is made up of two parts: a discriminator, which learns from incorrect input, and a generator, which learns to make false data.
GANs have become increasingly popular over time. They can be used in dark-matter investigations to simulate gravitational lensing and improve astronomical images. Using image training, video game developers may reproduce low-resolution, 2D visuals in earlier games in 4K or higher resolutions using GANs.
GANs are used to create 3D objects, generate realistic images and cartoon characters, and photograph people's faces.

Radial Basis Function Networks (RBFNs)

Radial basis functions (RBFs) are a special type of feed forward neural network that uses them as activation functions. These are composed of three layers: input, hidden, and output. Their principal applications include time-series prediction, regression, and classification.

Summary of differences: MACHINE-LEARNING vs. deep learning [21]

TABLE 1 SUMMARY OF ML VS DL

	Machine Learning	**Deep Learning**
What is it?	ML is an artificial intelligence (AI) methodology. Not all ML is deep learning.	Deep learning is an advanced ML methodology. All deep learning is ML.
Best suited for	ML is best for well-defined tasks with structured and labeled data.	Deep learning is best for complex tasks that require machines to make sense of unstructured data.
Problem solving approach	ML solves problems through statistics and mathematics.	Deep learning combines statistics and mathematics with neural network architecture.
Training	You have to manually select and extract features from raw data and assign weights to train an ML model.	Deep learning models can self-learn using feedback from known errors.
Resources required	ML is less complex and has a lower data volume.	Deep learning is more complex with a very high data volume.

From the table we conclude that both Machine Learning and Deep Learning are Artificial Intelligence technologies that may be used to analyze patterns, make predictions, and take actions on massive amounts of data. While they are connected, they are not the same thing. They differ in critical ways, such as how they learn and how much human assistance is required. Moreover, Machine Learning and Deep Learning are similar in that they both employ computers to identify and analyze data before making predictions. The main points of distinction are how they do it and what is expected of the individuals who develop it.

Machine Learning (ML) and Deep Learning (DL) are two subfields of Artificial Intelligence. AI is a subset of Machine Learning, and Deep Learning is a subset of ML (in other words, all Deep Learning is ML, but not all ML is Deep Learning).

2.3. Related Work

[23]Machine learning approaches were fine-tuned to learn the unlawful transaction patterns utilizing supervised algorithms such as logistic regression (LR), random forest (RF), and support vector machine (SVM).

[24]Supervised learning is good at identifying known manipulation kinds and variants. It uses manipulative data labeling, requiring previous knowledge of the manipulation type.

[25]The current strategy in industry for detecting market manipulation is top-down, relying on a collection of recognized patterns and predetermined criteria. Market data such as the price and volume of securities (i.e. the number of shares or contracts exchanged in a security) are tracked using a set of rules, and red flags trigger notifications.

[26] When implementing AI/machine learning approaches to trade surveillance, two essential dimensions must be considered: data processing and alert production. There are numerous use cases for each dimension in which machine learning might provide an economic advantage to financial institutions, depending on the business model and monitoring approach. In this paper, the author stresses the application of machine learning for trade monitoring and warning creation using a rule-based method. The author also seeks to utilize a case study and a road map to develop an AI-based off-the-shelf trade surveillance system that will free up surveillance and allow legal officers to focus on their core responsibilities.

[15] In this article, the writers attempt to define an anomaly on their own before reviewing a piece of writing on anomaly identification. The authors go on to discuss an article on how to tie "ML" algorithms to trade surveillance systems and how to use them to discover anomalies. The authors also distinguish between important price-based trade manipulations types, such as washed trade, spoofing, and others. Further, the author discusses the significance of having a dataset for having a good, scalable surveillance system as well as the application of rule-based surveillance giving such dataset that system gives alert to surveillance office but this alert may be a false-positive and a continuous calibration on the dataset must be applied to the dataset to come to true-positive results.

[27] The current strategy in industry for detecting market manipulation is top-down, relying on a collection of recognized patterns and predetermined criteria. Market data such as the price and volume of securities (i.e. the number of shares or contracts exchanged in a security) are tracked using a set of rules, and red flags trigger notifications.

[28] As computational finance technology advances, bad actors are given strong incentives to improve and create new instruments to influence markets. Humans using algorithms for illegal intentions is not a new phenomenon in banking. Given that AI may assist investment businesses in optimizing their company operations, transferring financial trading decision-making to AI systems may result in efficient algorithmic manipulation tactics and highly successful trading solutions.

[29] We employ supervised machine learning algorithms to identify market manipulation in China using information published by the China Securities Regulation Commission (CSRC) and data from the security market. In supervised machine learning, we primarily employ classification algorithms to discover anomalies in daily and tick trading data of manipulated stocks.
[30] For the first time, NASDAQ is using artificial intelligence on its US stock exchange to detect unusual and potentially harmful trading behavior. The newly formed effort intends to improve and modernize market monitoring by leveraging machine learning and other artificial intelligence (AI) technologies.

2.4. Summary

The Ethiopia Commodity Exchange now manages trade manipulation systems using a standard approach for both pre- and post-transaction. As previously said, the surveillance department thoroughly investigates pre-defined alterations to classify them as manipulation. This process can be time-consuming, and there is a chance that false positive findings will emerge. Thus, we put out a paradigm for self-aware AI base trade manipulation model.

There are several gaps to fill in order to complete an AI-based trade manipulation model, such as the accessibility of labeled data, attribute selection for executing the model, selecting the type of model to run, i.e. ML or DP, which evaluation matrix would best match for the model, and so on. These are the primary processes in obtaining the described surveillance model.

This research paper designs a machine learning-based trade surveillance model for the Ethiopian Commodity Exchange. The labeled data and the feature selection will be generated using Ethiopian commodity exchange trade data and historical market data.

This research paper will have the ability to predict trade manipulation base on the data that we get from trade data and historical market data. The model will be put on the Ethiopian Commodity Exchange monitoring department once it has been trained and evaluated.

CHAPTER TREE

3. Methodology

3.1. Data Collection and Preparation

There are numerous essential phases involved in collecting and processing data for the development of a model to identify market manipulation on the Ethiopian Commodity Exchange (ECX). This includes discovering relevant data sources, gathering information, cleaning and converting it, and preparing it for machine learning. The complete stages and considerations for this approach are provided below.

Manipulated trades were tagged based on data from the "ECX" manipulation rules. This chapter will examine the modeling and prediction of two manipulation activities, namely "Spoofing" using the order book data that we have specified below

TABLE 2 SAMPLE DATASET

Order Id	Order Quantity	Price	Symbol	Order time	Order Status	Member ID	IsClient Order	Order Type
B-M22076-308	8	10350	WHGS2	00:00.0	Expired	M22076	No	Buy
B-M22076-308	8	10750	WHGS2	00:00.0	Expired	M22076	No	Buy
B-M22076-308	8	10750	WHGS2	00:00.0	Expired	M22076	No	Buy
S-C1111255-23	2	11890	WHGS2	00:01.9	Accepted	M7743	Yes	Sell

S-C1111255-23	2	11890	WHGS2	00:01.9	Accepted	M7743	Yes	Sell
S-C1111255-23	2	11890	WHGS2	00:01.9	Accepted	M7743	Yes	Sell
S-C1111255-23	2	11950	WHGS2	00:01.9	Accepted	M7743	Yes	Sell
S-C1111255-23	2	11950	WHGS2	00:01.9	Accepted	M7743	Yes	Sell
S-C1111255-23	2	11950	WHGS2	00:01.9	Accepted	M7743	Yes	Sell
S-C1111255-23	2	11890	WHGS2	00:01.9	Accepted	M7743	Yes	Sell

The primary sources of data for the study paper are going to be historical trade data, order book and market data. Feature attributes like "OrderId," "OrderQty," "Price," "Symbol," "OrderTime," "OrderStatus," "MemberId," “OrderType (side),” and “IsClientOrder,” are contained in this data. We utilize 70,401 total data points, of which 15,000 are used for test data sets and the remaining amount can be used for training data sets.

3.2. Data Preprocessing

The process of transforming data from one form to another, so that it may be used more effectively and is more useful, is known as data processing. This entire process may be automated with the use of machine learning algorithms, statistical expertise, and mathematical modeling.

There are steps that we follow to process the data

Data Collection: as we mentioned on data collection the primary sources of data for the study paper are going to be historical trade and market data from the ECX. Feature attributes like

"OrderId," "OrderQty," "Price," "Symbol," "OrderTime," "OrderStatus," "MemberId," "OrderType (side)," and "IsClientOrder," are contained in this data.

Data preprocessing: The data must be cleaned, filtered, and transformed at this stage in order to prepare it for additional analysis. This might need transforming the data to a new format, scaling or normalizing the data, or eliminating missing numbers or removing NaN (Not a Number) and outlier data.

At this point, we eliminate irrelevant data to find the data that will be our feature selection. This data will be utilized for the research. "OrderType (side)," "OrderTime," "OrderStatus," "MemberId," "OrderId," "OrderQty," "Price," "Symbol," and "IsClientOrder."Data analysis: This phase entails analyzing the data using a range of techniques, such as statistical analysis, machine learning algorithms, and data visualization. Extracting knowledge or insights from the data that we preprocessed during the preprocessing stage is the goal of this phase.

Data interpretation: In this stage, the data analysis findings are interpreted, and conclusions are made using the newfound knowledge. Additionally, it could entail briefly and clearly presenting the results via dashboards, reports, or other visualizations.

Data storage and management: Following processing and analysis, the data has to be handled and stored in a safe and convenient manner. This might entail backing up and recovering the data to prevent loss, as well as storing it in a database, cloud storage, or other systems.

Data visualization and reporting: Finally, stakeholders are given access to the data analysis findings in a manner that is clear and practical. This might entail producing dashboards, reports, or visualizations that draw attention to important data patterns and discoveries.

3.2.1. Data Description

Our dataset is made up of "ECX" market data and order book, thus each piece of data has its own bid and ask pricing. This leads to a broad range of prices and inconsistent data. Bid and ask price is the best potential price that buyers and sellers are willing to execute a deal

As a result, all bid/ask prices were normalized in order to maintain consistency in the data while still identifying manipulative trades. Our full data set consists of nine columns of tick data (Bid/Bid Size, Ask/Ask Size), totaling approximately 278.62 K data points per column. These cases focused

on spoofing manipulation trades, with 70,401 bid/bid size quotes and with the suspected manipulated ask/ask size trades.

TABLE 3 DATA DESCRIPTION

IsClientOrder_num	Show the trade order is for client or member
OrderStatus_num	Show the status of the trade order “Accepted”, “Canceled”, “Rejected”
OrderType_num	Shows the order is “Buy order” or “Sell order”
OrderQuantity_num	Show the quantity given to the order
Price_num	Show the price given to the order
Ordertime_num	Show the time where the order submitted
MemberId_num	The Id given to the ECX member to trade

3.2.2. Feature Extraction and Feature Selection

Feature extraction and selection are critical phases in building a machine-learning model to identify market manipulation on the Ethiopian Commodity Exchange (ECX). These methods assist select and use the most relevant and informative information to improve the model's performance.

As previously said, feature extraction allows us to turn our data into features that more accurately depict the underlying problems. As a result, this study uses several graph-processing techniques, notably the continuous wavelet transform, to extract features.

Feature selection is a dimensionality reduction strategy that seeks to choose a small subset of useful characteristics from the original set by deleting unnecessary, redundant, or noisy features.

Feature selection typically results in improved learning performance, which includes higher learning accuracy, reduced computing cost, and better model interpretability. In general, irrelevant characteristics are those that cannot be used to differentiate between samples from various classes (supervised) or clusters (unsupervised). Irrelevant characteristics can be removed without affecting learning performance. In fact, removing unimportant characteristics may aid in the development

of a better model, as irrelevant information might confound the learning system and create memory and computation inefficiencies [31].

It is commonly recognized that bunch of features could end up in less accurate prediction analysis than less features. On the other side, feature selection is the process of identifying which features are most significant out of all of them. The domain expert specifies the independent and dependent variables in feature selection that have a positive or negative correlation with one another in order to forecast a model. Feature selection is a common way to minimize the problem of excessive and irrelevant features [32].

Another way of dimensionality reduction technique is feature extraction it is the process of obtaining additional features from two major features in order to improve the relevance and accuracy of the prediction model.

As previously stated, feature extraction involves creating new features from raw data that might aid in better catching patterns suggestive of market manipulation.

We extract the features using different parameters

Date and Time: Take out the month, day of the week, hour, and other time-related elements.

Features of lag: To represent temporal dependencies, provide lag versions of key characteristics like price and volume.

Statistics: Compute rolling means, variances, and other statistics over a window of time.

Price Changes: Compute the percentage change in prices.

Volume Spikes: Identify significant changes in trade volumes.

Volatility: Calculate the volatility of prices over a specified window.

Order Imbalance: Calculate the difference between buy and sell orders.

Order Flow: Measure the flow of orders over time.

Trader Activity: Measure the activity levels of traders.

Anomalous Behavior: Identify unusual trading patterns by specific traders.

Using the above parameters. We reduced this dimension to seven features data columns. This seven characteristic will decide the model that we intend to implement. Furthermore, the seven fields must be converted to integer values in order for the model to make predictions accurately. We will see all this in the implementation part of this research work in detail.

TABLE 4 FEATURE SELECTION

IsClient Order _num	Order Status _num	Order Type _num	Order Quantity _num	Price _num	Order time _num	Member Id _num	suspicious
0	0	0	0.120954	0.19649	0.13504	0.04819	0
0	0	0	0.120954	0.20491	0.13504	0.04819	0
0	0	0	0.120954	0.22597	0.13504	0.04819	0
0	0	0	0.120954	0.28842	0.13504	0.04819	0
0	0	0	0.120954	0.41965	0.13504	0.04819	0
...	...	...	...	...	...	...	...
0	0	1	0.875639	0.13614	0.72822	0.99398	1
0	0	1	0.875639	0.13614	0.72822	0.99398	1
0	0	1	0.875639	0.13614	0.72822	0.99398	1
0	0	1	0.875639	0.13895	0.72822	0.99398	1
0	0	1	0.875639	0.14105	0.72822	0.99398	1

3.3. Development Model

A number of crucial processes are involved in creating a machine-learning model to identify market manipulation on the Ethiopian Commodity Exchange (ECX): feature engineering, data preprocessing, model selection, training, assessment, and deployment.

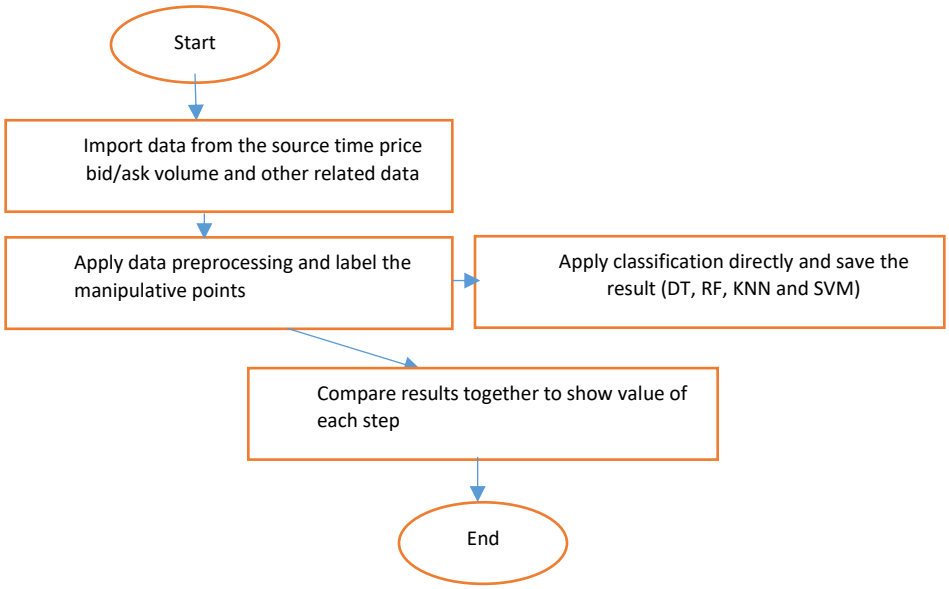

FIGURE 12 DEVELOPMENT MODEL

This study's main contribution is in the feature extraction segment of our workflow, however comparing different detection (Binary classification) Machine learning models can also be beneficial.

We use k-fold cross-validation to evaluate predictive models by partitioning the daily trading data and tick trading data into a training set to train the models, and testing set to evaluate them. For the supervised machine learning models, we use Random Forest Decision Tree (RFDT), K-nearest neighbors (KNN), Decision tree classifier (DTC), Logistic regression (LR), Gradient Boosting Machine (GBM) and Support Vector Machine (SVM) to build models.
Let us examine each one of them individually in this research study. We model the data using both supervised and unsupervised machine learning algorithms.

K-nearest neighbors (KNN)

A straightforward yet effective classification system called K-Nearest Neighbors (K-NN) may be used to identify instances of market manipulation on the Ethiopian Commodity Exchange (ECX). This comprehensive guide covers all the steps involved in implementing a K-NN model for this purpose, including feature engineering, training, evaluation, and cross-validation, as well as data pretreatment.

Random Forest Decision Tree

Another flexible supervised machine learning method that may be applied to regression and classification is called random forest. A collection of uncorrelated decision trees is referred to as a "forest" when they are merged in order to reduce variation and generate more precise data forecasts.

Decision tree classifier (DTC)

DTC is a non-parametric supervised learning approach for regression and classification. The objective is to build a model that can predict the value of a target variable based on fundamental decision rules derived from data properties. A piecewise constant approximation can be seen as a tree.

Logistic regression (LR)

Logistic regression is used when the dependent variable is categorical, meaning it contains binary outputs like "true" or "false" or "yes" or "no," as opposed to linear regression, which is used when the dependent variable is continuous. While finding correlations between data inputs is the goal of both regression techniques, logistic regression is mostly used to address binary classification problems such as spam detection.

Support vector machine (SVM)

The ideas of SVM are rather straightforward, and it is an intriguing method. Using a hyperplane with the greatest margin, the classifier divides data points. Because of this, another name for an SVM classifier is a discriminative classifier. In order to assist in the classification of new data points, SVM finds an ideal hyperplane.

3.4. Performance Evaluation

The concept of constructing machine learning, artificial intelligence, or deep learning models is based on the constructive feedback principle. You create a model, receive input from metrics, make modifications, and repeat until you achieve the desired classification accuracy. The evaluation measures explain the model's performance. A key component of assessment measures is their capacity to differentiate among model outputs [22].

The type of model and its implementation strategy are the only factors that influence the evaluation metric selection. Once the model creation process is complete, evaluation metrics will assist you in assessing the correctness of your model.

There are many evaluation metrics that we can apply on this research work here are some evaluation metrics that we apply.

Classification Accuracy

Arrangement when we use the term accuracy, we typically mean accurately. It is the ratio of the total number of input samples to the number of accurate predictions.

$$Accuracy = \frac{number\ of\ correct\ prediction}{total\ number\ of\ prediction\ made}$$

It works well only if there are equal number of samples belonging to each class.

Confusion Matrix

Confusion Matrix, as the name implies, returns a matrix that summarizes the model's overall performance.

Assume that we have a binary classification issue. We have several samples in two categories: YES or NO. We also have our own classifier that predicts a class based on a particular input sample. After testing our model on 165 samples, we obtained the following results.

TABLE 5 CONFUSION MATRIX

N=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

There are four important terms:

True Positives: The cases in which we predicted YES and the actual output was YES.

True Negatives: The cases in which we predicted NO and the actual output was NO.

False Positives: The cases in which we predicted YES and the actual output was NO.

False Negatives: The cases in which we predicted NO and the actual output was YES.

Accuracy for the matrix can be calculated by taking average of the values lying across the "main diagonal" i.e.

$$Accuracy = \frac{TruePositive + TrueNegative}{TotalSample}$$

$$i.e. Accuracy = \frac{100 + 50}{165} = 0.91$$

Logarithmic Loss

Logarithmic Loss, or "Log Loss," penalizes erroneous classifications. It performs well in multi-class classification. When working with Log Loss, the classifier must assign probability to each class over all data. Assuming there are N samples belonging to M classes, the Log Loss is calculated as follows:

$$LogLoss = \frac{-1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} yij * \log(pij)$$

Where,

yij, indicates whether sample i belongs to class j or not

pij, indicates the probability of sample i belonging to class j

Log Loss has no upper bound and it exists on the range [0, ∞). Log Loss nearer to 0 indicates higher accuracy, whereas if the Log Loss is away from 0 then it indicates lower accuracy.

In general, minimizing Log Loss gives greater accuracy for the classifier.

F1 Score

F1 Score is the Harmonic Mean of accuracy and recall. The range of F1 Score is [0, 1]. It indicates your classifier's precision (the number of cases it properly classifies) and robustness (the number of instances it does not miss).High accuracy but poor recall produces incredibly precise results, but it also misses a huge number of occurrences that are difficult to identify. Our model's performance improves as the F1 Score increases. It may be mathematically represented as follows:

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

F1 Score tries to find the balance between precision and recall.

Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$precision = \frac{TruePsitive}{TruePositive + FalsePositive}$$

Recall: It is the number of accurate positive findings divided by the total number of relevant samples (all samples that should have been classified as positive).

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Mean Absolute Error

The Mean Absolute Error (MAE) is the average of the difference between the original and predicted values. It tells us how distant the forecasts were from the actual results. However, they do not indicate the direction of the inaccuracy, i.e., whether we are underestimating or overestimating the data. The mathematical representation is as follows:

$$MeanAbsoluteError = \frac{1}{N}\sum_{j=1}^{N} |yi - \hat{y}j|$$

Mean Squared Error

Mean Squared Error (MSE) is identical to Mean Absolute Error; the main difference is that MSE calculates the average of the square of the difference between the actual and forecasted values. MSE has the benefit of being easier to compute the gradient than Mean Absolute Error, which needs complex linear programming techniques. As we square the error, the influence of greater mistakes becomes more evident than smaller ones, allowing the model to focus more on the larger faults.

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(yi - \hat{y}j)^2$$

3.5. Proposed Architecture

This architecture offers a complete and scalable framework for creating, implementing, and administering a market manipulation detection system for the Ethiopian Commodity Exchange that employs machine learning.

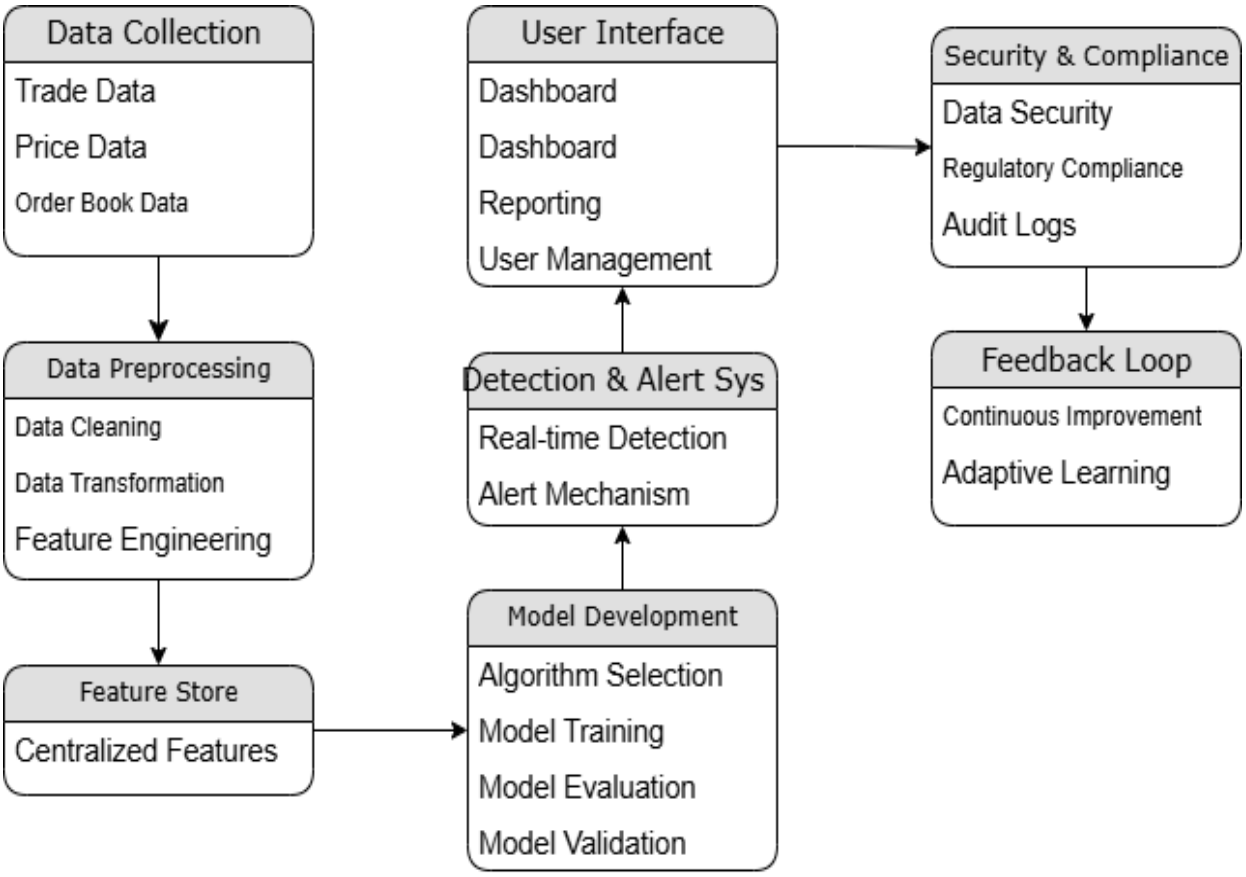

FIGURE 13 PROPOSED ARCHITECTURE

CHAPTER FOUR

4. Design/Implementation/Experimental Results/Discussions

This chapter outlines the detailed implementation steps for developing the machine-learning model to detect market manipulation on the ECX.

Sources of Data

- Historical Price Information: Taken from the ECX database, encompassing opening, high, low, and daily closing values.
- Trade Volumes (Refer to Quantity): Daily trade volumes of various commodities according to quantity volume.
- Order Book Data: Information on the buy and sell orders, including quantities and prices.
- Transaction Records: Detailed logs of completed trades.

Data Extraction

Pandas Data Reader is a stand-alone tool that extracts data from many web sources and stores it in pandas DataFrames. It is a useful tool for financial research and other applications that require data from web sources.

Data Preprocessing

Data preparation is an important step before beginning to analyze or model the ECX datasets. It guarantees that the data is clean and consistent, allowing machine learning algorithms or statistical tests to give accurate findings. Here is an overview of popular data preparation approaches for financial data.

Handling Missing Values

The ECX market data or order book data may have missing numbers for a variety of causes. Techniques for addressing them include.

- **Deletion:** If there are few missing data, points and they are unlikely to cause bias, eliminating rows or columns with missing values may be appropriate.

- **Charging (Imputation):** Filling in missing values with estimates. Common methods include:
    - **Mean/Median/Mode:** Replace missing values with the average, median, or most frequent value in the column.
    - **Interpolation:** Estimate missing values based on surrounding data points. Techniques like linear interpolation or forward/backward fill can be used.
- **Model-based methods:** Use statistical models to predict missing values based on other features in the data.

Dealing with Outliers

Outliers are data points that deviate significantly from the rest. They can distort analysis results. Here are some approaches:

- **Trimming:** Remove a small percentage of extreme outliers from either end of the distribution.
- **Investigate:** If outliers are legitimate, consider maintaining them after determining the underlying causes.

Feature Engineering

- Create new features that might be more informative for your analysis. This could involve:
    - Scikit-Learn's MinMaxScaler is a preprocessing technique that scales and modifies dataset characteristics to a defined range, often 0 to 1. This scaling is especially important for machine learning algorithms that require features to have comparable ranges in order to avoid specific characteristics from dominating the process. MinMaxScaler can help models perform better and converge faster.

Normalization or Standardization

- Features in Ethiopian Commodity Exchange or any financial datasets frequently have varying scales. Normalization or standardization helps to align them on a common scale, which improves model performance.
    - **Normalization:** Scales features to a range between 0 and 1 (MinMax Scaling).

- **Standardization:** Scales features to have a mean of 0 and a standard deviation of 1 (Z-score normalization).

```
cols_to_scale = ['IsClientOrder_num','OrderStatus_num','OrderType_num','OrderQuantity_num','Price_num','Ordertime_num','MemberId_num','suspicious']

from sklearn.preprocessing import MinMaxScaler

    scaler = MinMaxScaler()
    data_frame[cols_to_scale] = scaler.fit_transform(data_frame[cols_to_scale])
```

As we can see from the code sample, we utilize MinMaxScaler() from the sklearn package and adjust all the data values to scale as integral numbers.

TABLE 6 SCALE THE DATA TO INTEGRAL NUMBER

IsClient Order _num	Order Status _num	OrderType _num	Order Quantity _num	Price _num	Order Time _num	Member Id _num	suspicious
1.0	0.0	1.0	0.490385	0.661130	0.000000	0.924242	0.0
1.0	0.0	1.0	0.490385	0.661130	0.000000	0.924242	0.0
1.0	0.0	1.0	0.490385	0.661130	0.000000	0.924242	0.0
1.0	0.0	1.0	0.490385	0.707641	0.000000	0.924242	0.0
1.0	0.0	1.0	0.490385	0.707641	0.000000	0.924242	0.0
...	...	...	...	...	...	...	...
0.0	1.0	0.0	0.990385	0.438538	0.995261	0.787879	1.0
0.0	1.0	0.0	0.990385	0.438538	0.995261	0.787879	1.0
0.0	1.0	0.0	0.990385	0.438538	0.995261	0.787879	1.0
0.0	1.0	0.0	0.990385	0.438538	0.995261	0.787879	1.0
1.0	0.0	1.0	0.221154	0.448505	1.000000	0.000000	0.0

By processing the data using the above technique we identify the independent and dependent variable. From this all the column "data_frame.iloc[:, [0,1,2,3,4,7]].values" are the independent variable and "suspicious" is dependent variable.

4.1. Model Selection

We will consider the following models for comparison:

- Logistic Regression
- Decision Trees
- Random Forest
- Gradient Boosting Machines
- Support Vector Machines
- Neural Networks

4.1.1. Logistic Regression

This section outlines the development, training, and evaluation of a Logistic Regression model for detecting market manipulation on the ECX.

Load and Preprocess Data: First, we need to load the necessary datasets and preprocess the data to make it suitable for model training.

Training the Logistic Regression Model: We will use the Logistic Regression model from the "sklearn" library. Logistic Regression is suitable for binary classification problems and provides probabilities for class membership.

Hyper-Parameter selection: Hyper-parameter tuning has considerably enhanced the Logistic Regression model's performance in identifying market manipulation on the ECX. The tuned model outperforms the default model in terms of accuracy, precision, recall, and F1 score. The insights collected from the feature significance analysis can aid in determining the primary elements that contribute to market manipulation.

```
class LogisticRegression(
penalty: Literal['l1', 'l2', 'elasticnet'] | None = "l2",
*,
dual: bool = False,
tol: Float = 0.0001,
C: Float = 1,
fit_intercept: bool = True,
```

intercept_scaling: Float = 1,

class_weight: Mapping | str | None = None

)

Define Labels: Establish the labels for the logistic regression model, such as (1) and (0), indicating whether a transaction is regarded as manipulative. This might entail utilizing known manipulation examples for supervised learning or expert labeling. As we show on the above the column "**suspicious**"

Model Building: With the aid of the abovementioned hyper-parameter, we can construct the logistic regression model using a Python-based methodology and libraries like those that pandas, scikit-learn, and numpy.

Model Evaluation: Use measures such as F1-score, recall, accuracy, and precision to assess the model. The model's ability to identify tampering is demonstrated by the confusion matrix and classification report. See the table for more information

TABLE 7 EVALUATION MATRIX FOR LOGISTIC REGRESSION MODEL

	precision	recall	f1-score	support
0	0.86	0.94	0.9	282
1	0.84	0.68	0.75	130
accuracy			0.86	412
macro avg	0.85	0.81	0.82	412
weighted avg	0.86	0.86	0.85	412

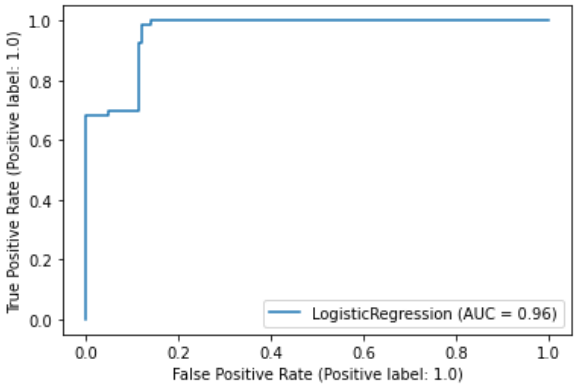

FIGURE 14 LOGISTIC REGRESSION ACCURACY GRAPH

4.1.2. Decision Trees

The model will be evaluated using strong performance indicators such as accuracy, precision, recall, and the F1 score. These metrics will assess how successfully the model recognizes potential cases of market manipulation. The results aim to increase openness and confidence in commodities trading in Ethiopia's rural economy by providing helpful information to ECX authorities and stakeholders in order to improve supervision and regulatory measures. This study contributes to the larger goal of protecting market integrity and fostering fair trading practices in Ethiopian commodity markets by providing an effective model for identifying ECX market manipulation.

Hyper-Parameter selection: Hyper-parameter tuning has considerably enhanced the decision tree model's performance in identifying market manipulation on the ECX. The tuned model outperforms the default model in terms of accuracy, precision, recall, and F1 score. The insights collected from the feature significance analysis can aid in determining the primary elements that contribute to market manipulation.

Model Building: With the aid of the hyper-parameter, we can construct the decision tree model using a Python-based methodology and libraries like those that pandas, scikit-learn, and numpy.

Model Evaluation: Use measures such as F1-score, recall, accuracy, and precision to assess the model. The model's ability to identify tampering is demonstrated by the confusion matrix and classification report. See the table for more information

Table 8 EVALUATION MATRIX FOR DECISION TREE MODEL

	precision	recall	f1-score	support
0	1	0.78	0.88	282
1	0.68	1	0.81	130
accuracy			0.85	412
macro avg	0.84	0.89	0.84	412
weighted avg	0.9	0.85	0.86	412

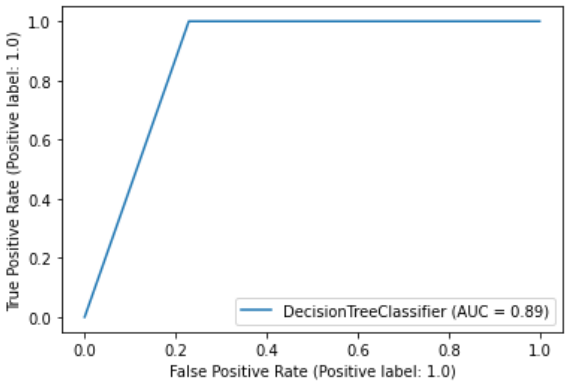

FIGURE 15 DECISION TREE MODEL ACCURACY GRAPH

4.1.3. Random Forest

Developing a Random Forest model to identify market manipulation on the Ethiopian Commodity Exchange (ECX) requires multiple phases. This strategy uses ensemble learning to increase prediction accuracy. A full implementation of the model is provided below, along with an example implementation in Python and Scikit-learn. All the step are the same with logical regression model that we mention above.

Hyper-Parameter selection: Hyper-parameter tuning is crucial for optimizing the performance of a Random Forest model. The most common hyper-parameters to tune in a Random Forest model include:

- n_estimators: The number of trees in the forest.
- max_depth: The maximum depth of the tree.

- min_samples_split: The minimum number of samples required to split an internal node.
- min_samples_leaf: The minimum number of samples required to be at a leaf node.
- max_features: The number of features to consider when looking for the best split.

Model Building: With the aid of the hyper-parameter, we can construct the random forest tree model using a Python-based methodology and libraries like those that pandas, scikit-learn, and numpy.

Model Evaluation: Use measures such as F1-score, recall, accuracy, and precision to assess the model. The model's ability to identify tampering is demonstrated by the confusion matrix and classification report. See the table for more information.

TABLE 9 RANDOM FOREST EVALUATION MATRIX

	precision	recall	f1-score	support
0	0.86	0.91	0.89	282
1	0.79	0.68	0.73	130
accuracy			0.84	412
macro avg	0.82	0.8	0.81	412
weighted avg	0.84	0.84	0.84	412

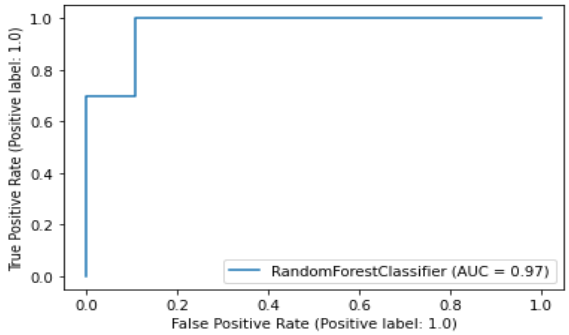

FIGURE 16 RF ACCURACY GRAPH

4.1.4. Gradient Boosting Machines

Similar steps to those in the Random Forest model must be followed in order to create a Gradient Boosting Machine (GBM) model for identifying market manipulation on the Ethiopian Commodity Exchange (ECX). This model makes use of the gradient boosting algorithm, which builds an ensemble of trees sequentially, with each tree trying to correct the errors of the previous one.

Hyper-Parameter selection: Hyper-parameter tuning is crucial for optimizing the performance of a Gradient Boost Machine. The most common hyper-parameters to tune in a Gradient Boost Machine model include:

- n_estimators: The number of trees in the forest.
- max_depth: The maximum depth of the tree.

- min_samples_split: The minimum number of samples required to split an internal node.
- min_samples_leaf: The minimum number of samples required to be at a leaf node.
- max_features: The number of features to consider when looking for the best split.

Model Building: With the aid of the hyper-parameter, we can construct the GBM model using a Python-based methodology and libraries like those that pandas, scikit-learn, and numpy.

Model Evaluation: Use measures such as F1-score, recall, accuracy, and precision to assess the model. The model's ability to identify tampering is demonstrated by the confusion matrix and classification report. See the table for more information

Table 10 GMB ACCURACY MATRIX

	precision	recall	f1-score	support
0	0.88	1	0.93	282
1	1	0.69	0.82	130
accuracy			0.9	412
macro avg	0.94	0.85	0.88	412
weighted avg	0.91	0.9	0.9	412

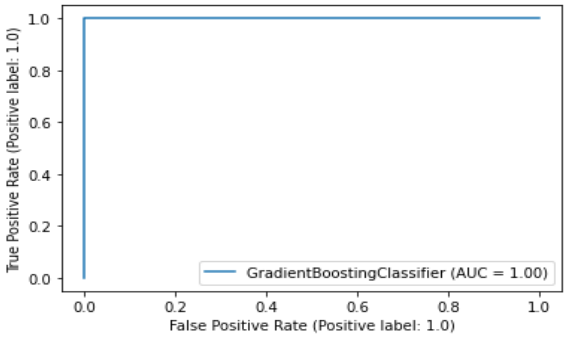

FIGURE 17 GMB ACCURACY GRAPH

4.1.5. Support Vector Machines

Similar procedures to those for creating other machine learning models must be followed in order to create a Support Vector Machine (SVM) model for identifying market manipulation on the Ethiopian Commodity Exchange (ECX). This model makes use of the SVM algorithm, which works especially well in high-dimensional spaces and situations where there are more dimensions than samples.

Hyper-parameter finding the ideal settings for the Support Vector Machine (SVM) that yield the highest model performance requires hyper-parameter adjustment. The most crucial SVM model hyper-parameters to adjust are:

- C: Regularization parameter. It controls the trade-off between achieving a low error on the training data and minimizing the complexity of the model.
- kernel: Specifies the kernel type to be used in the algorithm. Common kernels include 'linear', 'poly', 'rbf', and 'sigmoid'.

- gamma: Kernel coefficient for ‘rbf’, ‘poly’, and ‘sigmoid’. It defines how far the influence of a single training example reaches.
- degree: Degree of the polynomial kernel function (‘poly’). It's relevant only when the kernel is 'poly'.

Model Building: With the aid of the hyper-parameter, we can construct the SVM model using a Python-based methodology and libraries like those that pandas, scikit-learn, and numpy.

Model Evaluation: Use measures such as F1-score, recall, accuracy, and precision to assess the model. The model's ability to identify tampering is demonstrated by the confusion matrix and classification report. See the table for more information

TABLE 11 SVM ACCURACY MATRIX

	precision	recall	f1-score	support
0	0.86	0.93	0.9	282
1	0.82	0.68	0.74	130
accuracy			0.85	412
macro avg	0.84	0.8	0.82	412
weighted avg	0.85	0.85	0.85	412

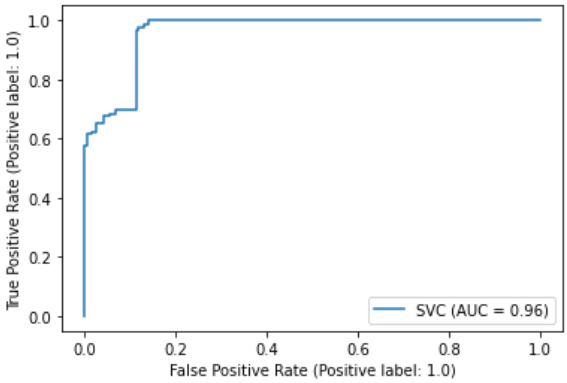

FIGURE 18 SVM ACCURACY GRAPH

CHAPTER FIVE

5. Conclusions and Future Works

Using a variety of machine-learning algorithms—each with pros and cons—to develop a model for identifying market manipulation on the Ethiopian Commodity Exchange (ECX) is necessary.

Although decision trees offer a simple and easy-to-understand model, its propensity to over fit need careful tweaking and maybe group techniques in order to maximize performance. With hyper-parameter adjustment, Random Forest is an even more potent contender for identifying market manipulation since it provides a solid mix between robustness and accuracy. GBM is a potent model with great accuracy for identifying market manipulation. Although it has to be carefully adjusted, it can successfully identify intricate patterns in the data. SVM is a reliable option for identifying market manipulation, particularly in high-dimensional domains. For best results, proper hyper-parameter adjustment is essential.

Model Selection: Based on the model complexity and simples we select the model accordingly.

- **Easy models**: Due to their ease of use and interpretability, start with decision trees and logistic regression.
- **Advanced Models**: For increased accuracy and resilience, switch to ensemble techniques like Random Forest and Gradient Boosting.
- **Specialized Models**: SVM should be used in certain situations involving high-dimensional feature spaces and reasonably sized data sets.

Performance and Trade-offs: Based on trade-off and there performance we categorize as below

- **Trade-off**: Complexity versus Simplicity. While simpler models like decision trees and logistic regression are simpler to understand, they might not be as accurate. Better performance is provided by more intricate models (Random Forest, GBM, SVM), but they also need more computing power and meticulous tweaking.

- **Accuracy**: The greatest results are often obtained by ensemble approaches (Random Forest, GBM) for difficult problems like detecting market manipulation.

Implementation:

- **Hyper-parameter Tuning**: To improve model hyper-parameter performance, use Grid Search or Random Search with Cross-Validation.

Future Work:

Feature engineering: Constantly improve and expand features that grasp the subtleties of manipulating the market.
Model Updating: To adjust to changing manipulation techniques, update the model on a regular basis using fresh data.
A reliable and efficient method for identifying market manipulation on the ECX may be created by carefully choosing and fine-tuning these machine-learning models, greatly enhancing the integrity and fairness of the market.

References

[1] W. Kento, "marketsurveillence.asp," investopedia, 2022. [Online]. Available: https://www.investopedia.com/terms/m/marketsurveillence.asp.

[2] D. Cumming and S. Johan, Global market surveillance, Tilburg University, 2008.

[3] X. S. X. e. a. Li, Design Theory for Market Surveillance Systems,

ocial Science Research Network (SSRN), 2022.

[4] K. Ayech, "AN ASSESSMENT ON MARKET INTEGRITY OF ETHIOPIAN," smu, 2019.

[5] W. &. Mary, "What Exactly is Market Integrity?," in *What Exactly is Market Integrity?*, William & Mary Law School Scholarship, 2017.

[6] E. C. Exchange, "AboutUs.aspx," 2015. [Online].

Available: https://www.ecx.com.et/Pages/AboutUs.aspx.

[7] A. J. A. O. H. J. O. O. O. J. Osisanwo F.Y., Supervised Machine Learning Algorithms: Classification and Comparison, International Journal of Computer Trends and Technology (IJCTT), 2017.

[8] D. Cumming and S. Johan, Global market surveillance, Tilburg University, 2008.

[9] C. Pirrong, Commodity Market Manipulation Law: A (Very) Critical Analysis, Wash. & Lee L. Rev., 1994.

[10] C. Pirrong, The Economics of Commodity Market, Bauer College of Business, 2017.

[11] IBM, "artificial-intelligence," [Online].

Available: https://www.ibm.com/topics/artificial-intelligence.

[12] B. Mahesh, Machine Learning Algorithms -A Review,

International Journal of Science and Research (IJSR), 2020.

[13] IBM, "machine-learning," 2023. [Online].

Available: https://www.ibm.com/topics/machine-learning.

[14] B. Mahesh, "Machine Learning Algorithms -A Review,"

in *Machine Learning Algorithms -A Review*, 2019, p. 381.

[15] H. R. H. L. Shweta Tiwari, Machine Learning in Financial Market,

1Norwegian University of Science and Technology, 2021.

[16] G. Cloud, "Google Cloud," [Online]. Available:

https://cloud.google.com/discover/what-is-unsupervised-learning#:~:text=

Unsupervised%20learning%20algorithms%20are

%20better,features%20useful%20for%20categorizing%20data..

[17] IBM, "Naïve Bayes classifiers," 2023. [Online]. Available: https://www.ibm.com/topics/naive-bayes?mhsrc=ibmsearch_a&mhq=Naive%20bayes.

[18] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, Michael S. Lew,

"Deep learning for visual understanding," in *Deep learning for visual understanding*, 2015.

[19] IBM, "deep-learning," 2023. [Online]. Available: https://www.ibm.com/topics/deep-learning.

[20] IBM, "What is supervised learning?," 2023. [Online]. Available: https://www.ibm.com/topics/

supervised-learning#:~:text=the%20next%20step-,What%20is%20supervised%

20learning%3F,data%20or%20predict%

20outcomes%20accurately..

[21] Amazon, "the-difference-between-machine-learning-and-deep-learning," 2023. [Online]. Available: https://aws.amazon.com/compare/the-difference-between-machine-learning-and-deep-learning.

[22] T. Srivastava, "11-important-model-evaluation-error-metrics," 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/#:~:text=Evaluation%20metrics%20are%20quantitative%20measures,comparing%20different%20models%20or%20algorithms..

[23] M. E. A. F. M. Mohd Asyraf Zulkifley, Stock Market Manipulation Detection using Artificial Intelligence, DASA, 2021.

[24] P. T. Teema Leangarun, Stock Price Manipulation Detection Using Deep, IEEEAccess, 2021.

[25] O. R. Z. D. D. S.K. Golmohammadi, Detecting Stock Market Manipulation using Supervised Learning Algorithms, ResearchGet, 2015.

[26] P. Faulstich, Machine Learning in Trade Surveillance, l-p-a.com, 2023.

[27] O. R. Z. D. D. Koosha Golmohammadi, Detecting Stock Market Manipulation using.

[28] W. G. R. h. S. S. Alessio Azzutti, "MACHINE LEARNING, MARKET MANIPULATION,," in *MACHINE LEARNING, MARKET MANIPULATION,*, Penn Law:, p. 97.

[29] J. W. Z. L. Aihua Li, Market Manipulation Detection Based on Classification Methods, Central University of Finance and Economics, Beijing, 2017.

[30] Nasdaq, "for-the-first-time-nasdaq-is-using-artificial-intelligence-to-surveil-u.s.-stock-market," 2019. [Online]. Available: https://www.nasdaq.com/articles/for-the-first-time-nasdaq-is-using-artificial-intelligence-to-surveil-u.s.-stock-market.

[31] H. L. J. T. Suhang Wang, "Feature Selection," *Feature Selection*, p. 1, 2016.

[32] T. F. A. W. K.-L. J. M. O. Nicholas Pudjihartono, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction*, 2022.

[33] A. J. A. O. H. J. O. O. O. A. J. Osisanwo F.Y., "Introduction," in *Supervised Machine Learning Algorithms: Classification and Comparison*, IJCTT, 2017, p. 128.