American University in Cairo

AUC Knowledge Fountain

Theses and Dissertations

Student Research

Spring 6-15-2021

Stock Market Manipulation Detection Using Continuous Wavelet Transform & Machine Learning Classification

Sarah Youssef
s_moh15@aucegypt.edu

The American University in Cairo

School of Business

Stock Market Manipulation Detection Using Continuous Wavelet Transform & Machine Learning Classification

A Thesis Submitted to

Department of Management

in partial fulfillment of the requirements for the degree of

Master of Science in Finance

by

Sarah Mohamed Youssef

under the supervision

of Dr.Medhat Hussanien

Co-Supervised by Dr Noha Youssef

December 2020

Abstract

Stock market manipulation detection is important for both investors and regulators. Being able to detect stock manipulation and preventing it gives investors the confidence in the market fairness and integrity. It also helps maintaining liquidity of the stocks and market efficiency.

Implementing data mining algorithms in manipulation detection is a relatively recent technique but in the past few years there has been an increasing interest in it's applications in this domain. The benefit of monitoring manipulative trade behavior is that it can be implemented on live feed of stock data, which saves a lot of time in detecting stock price manipulation.

This research implements machine learning algorithms in detecting trade manipulations where trade behaviors artificially impact the National Best Bid and Offer (NBBO) of traded stocks. Research methodology implemented is based on feature extraction using signal analysis, taking advantage of the similarity between physical signals measured by machines and raw financial data. Accordingly, Continuous Wavelet Transform (CWT) is applied on actual manipulation data for feature extraction, Principal Component Analysis (PCA) and factor analysis are used for dimensionality reduction and then Machine Learning Classifiers are trained and tested. Tick Bid/Ask Price and volume data of actual 15 manipulation cases published by the Security Exchange Center (SEC) was extracted from an online interface and labeled accordingly. This data was then used to train, and test 3 different classification models (XGBoost, KNN & SVM) and the outcome was compared accordingly.

Results showed that introducing continuous wavelet transform enhances model accuracy, it increased precision results tremendously, while reducing recall values slightly. Adding PCA, reduced run time greatly, yet reduced the quality of some models prediction. Out of the three classifiers XGboost & KNN are showing the highest performance.

Contents

I Introduction 6

1 Introduction 6

2 Stock Market Manipulation 7

2.1 Definition 7

2.1.1 Background and Legal Definition 7

2.2 Stock Manipulation Types 7

II Literature Review 12

3 Stock Price Manipulation Detection 12

3.1 Input Variables 12

3.2 Feature Engineering/Dimensionality Reduction 13

3.2.1 What is a feature? 13

3.2.2 What is Feature engineering? 13

3.2.3 Why treat Financial data as Signal data? 15

3.2.4 Signal Processing: 15

3.3 Data Mining in Price Manipulation Detection (Detection model) 20

III Research Methodology 25

4 Detection System: 25

4.1 Proposed Work Flow 25

4.2 Data: 25

4.2.1 Data Source 25

4.2.2 Data Description 27

4.3 Feature Extraction 31

4.4 Dimension Reduction (PCA) 32

4.5 Detection Model 32

4.6 Evaluation Matrix 33

IV Results & Discussion 35

V Conclusion 43

List of Figures

1 Taxonamy of Manipulation Techniques by Putnins, Alexander, C & Cumming, D. (2020). . . . 9
2 Example illustrating Spoofing & Quote stuffing by Zhai, J., Zhai, J., Cao, Y., Cao, Y., Ding, X., & Ding, X. (2018). . . . 10
3 Layering Example from “What is the difference between layering and spoofing? (2017, June 12)” . . . 10
4 Feature Engineering in ML Workflow by Zheng, A., Casari, A., & Safari, an O{2019}[21] . . . 13
5 Feature Engineering break down in literature by Khalid, S., Khalil, T., & Nasreen, S. (2014). [3] . . . 14
6 Wavelet Review by Rhif, M., Ben Abbes, A., Farah, I., Martínez, B., & Sang, Y. (2019)[20] . . . 16
7 The Little Wave: (a) some wavelets, (b) location and (c) Scale by Addison, P. S. (2002)[15] . . . 17
8 The Wavelet, The signal and the Transform by Addison, P. S. (2002)[15] . . . 18
9 CWT Vs DWT from Wavelet Toolbox Matlabhelp . . . 19
10 Comparison of results from literature for WT in Industrial applications by Chattopadhyay, P., & Konar, P. (2014)[17] . . . 20
11 Flow Chart Shows Cases that will be tested . . . 25
12 Figure above shows Bid price offer and the size of the marker shows the bid size, The second plot shows the labeled data (0,1) 28
13 Actual Manipulation example from our dataset . . . 29
14 Actual Manipulation example from our dataset . . . 30
15 Demonstration of shrink and scaled Morl Wavelet, Scale is inversely proportional to frequency by Feike, S. (2020, February 10)[23] . . . 31
16 Scalogram Example applying scale 1-> 128 on our dataset Bid Prices . . . 32
17 Plot on top shows Confusion Matrix &Table below shows how each Metric can be calculated. Table by Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019) [2] . . . 34
18 Comparison Above showing 3 classifiers results for both Bid & Ask Data . . . 36
19 Case_2_Run_1_Bid Data Classifiers Comparison . . . 37
20 Case_2_Run_1_Ask Data Classifiers Comparison . . . 38
21 Case_2_Run_2_Bid Data Classifiers Comparison . . . 39
22 Case_2_Run_2_Ask Data Classifiers Comparison . . . 40
23 Case_3_Results both Bid & Ask Data . . . 41

Part I
Introduction

1 Introduction

Manipulation is any intentional action that is taken to interfere in the market mechanism impacting the security price weather by an increase or a decrease. However the lack of research for an effective and efficient detection model makes it hard for regulators to detect manipulative acts realtime.

In academia Allen and Gale were one of the first researchers to introduce manipulative trading. They divided it into three primary categories, information or action or trade-based manipulation. Information based manipulation is manipulation taking place by spreading wrong information that affects the security price. This can be linked to social media manipulation, where people can spread misleading information impacting the market [10].

Action based manipulation takes place when a person's actions can impact the supply and demand of a security accordingly impacting its price. Trade based manipulation does not contain any illegal activities instead it happens through legal trading activities of selling and buying offers or trades to impact the security prices. This includes spoof trading, quote stuffing, wash trades [10]& cornering the market.

The existing methodology in the industry to detect price manipulation or disruptive trading is a top down approach. It is based on already identified patterns, red flags or predetermined thresholds. Security data such as quotes for price and volumes are monitored using a set of boundaries and rules. When one of these rules are broken it generates an alert. This system is based on knowledge from experts but has two drawbacks, 1) might not be able to detect abnormal periods that are related to unknown manipulative schemes and 2) it might not be able to adapt to the fast changing market conditions as the amount of transactions is exponentially increasing (with the high frequency trading and increasing number of investors and listed securities).[10]

Data mining algorithms can be used to detect manipulation based on historical data (bottom-up approach). Golmohammadi, Zaiane, & Diaz, in 2015 identified 5 main categories according to which data mining can be applied to stock manipulation detection: 1- Social Network Analysis, 2- Visualization, 3- Rule Induction, 4- Outlier Detection and 5- Pattern Recognition using Supervised Learning methods. [10]

In our study we will be focusing on applying data mining techniques to detect anomalies or outliers in stock bid/ask prices and volumes. We will be proposing a novel approach for feature extraction, dimensionality reduction and run the output through different machine learning classifiers.

Our study is divided into 5 chapters. First chapter introduces different types of stock price manipulation and focuses more on spoofing and layering. Second chapter is literature review about different data mining techniques used by different authors in the same domain then we focus in the third chapter on our plied methodology. In the last two chapters we demonstrate the efficiency of the proposed technique and discuss its results and conclusion.

2 Stock Market Manipulation

2.1 Definition

2.1.1 Background and Legal Definition

Regulations to reform the securities market started since the market crash of 1929 as it was a widespread belief that it was a prime cause of the ensuing depression and that market manipulation and "excessive" speculations were behind the market crash. Similarly to 1929 crash [31], the damage brought by 2008 Recession generated the need for another reform, ultimately resulting in the Dodd–Frank Act. Before that, the SEC could prohibit individuals who breached either the Exchange Act or the Advisers Act from associating with various people in the securities world, including stockbrokers, dealers and investment advisers. The Dodd–Frank Act extended this authority as the commission can now also prohibit violators from associating with rating organizations or municipal advisers. [25]

According to 1934 Securities and Exchange Act (Exchange Act), the term "manipulation" may, be applied to any practice which has as its purpose the intentional raising, lowering or pegging of security prices[32]. In a free and open market it is a natural consequence of buying and selling that the price would be impacted yet once it is done intentionally and artificially it is considered manipulation[32]. Manipulated prices do not reflect the real supply and demand of the securities nor they reflect its liquidity instead it becomes misleading. The 1940 Investment Advisers Act (Advisers Act) prohibits almost the same behavior. The Act makes it illegal for ""any investment adviser" to "employ any device, scheme, or artifice to defraud any client or prospective client" or to "engage in any act, practice, or course of business which is fraudulent, deceptive, or manipulative."" 15 U.S.C. § 80b–6(1). [25]

According to the above definitions, to successfully claim price manipulation against someone under the SEA, the accuser must prove that "(1) the defendant has the ability to influence market prices; (2) an artificial price existed; (3) the defendant caused this artificial price; and (4) the defendant specifically intended to cause the artificial price." Specific intent means that this person has in purpose and consciously acted to impact a price in the market that is not a true reflection of the forces of supply and demand. By 1984, the SEC adapted the intent-based approach and in that year, the CFTC, Federal Reserve, and the SEC stated that intent was a vital element for all market manipulation claims. [33]

Therefore, rather than direct manipulation identification, this study tackles the detection of disruptive trading behaviors after which more investigation can be done by regulators to determine the intention behind it.

2.2 Stock Manipulation Types

The term "Market Manipulation" includes different manipulation strategies. Different authores tried to describe them and map them out.

In academia, Allen and Gale [1] were one of the first researchers to introduce manipulative trading. They divided it into three primary categories as mentioned earlier information-based, trade or action-based manipulation & trade-based manipulation. Information-based manipulation is manipulation taking place by spreading wrong information that affects the security price. This can be linked to social media manipulation, where people can spread misleading information impacting the market [10]. Action-based manipulation happens because of a person's actions that impacts the supply and demand of a security, accordingly impacting its price. Trade-based manipulation is not based on illegal actions instead it is applied trough lawful trading activities of buying and selling offers to impact the security prices. This includes spoof trading, quote stuffing, wash trades [10] & cornering the market.

One of the recent taxonomies published in 2020 was by C. Alexander and D. Cumming shown in Figure 2.1[17]. This taxonomy is an extension to previous work in 2012 by Putnins. The author grouped different strategies into Runs, Contract based, Spoofing/Order based and Market power. These categories are then divided based on the mechanisms that were used to faciliatate the manipulation such as Trade-based, Info-based, Action-based & Order-Based. Such strategies usually don't happen in isolation, one would expect some hypermodel that would take place.

This taxnomay shown in figure 1 is different from the earlier work by Allen and Gale in the distinguish between Trade-based and Order-based manipulation. Since the advancement in technology introduced High frequency trading, traders now can submit orders in large quantities and cancel them in milliseconds before they get executed. Accordingly this still creates the impact of manipulation and impacts securities prices without a trade actually taking place.

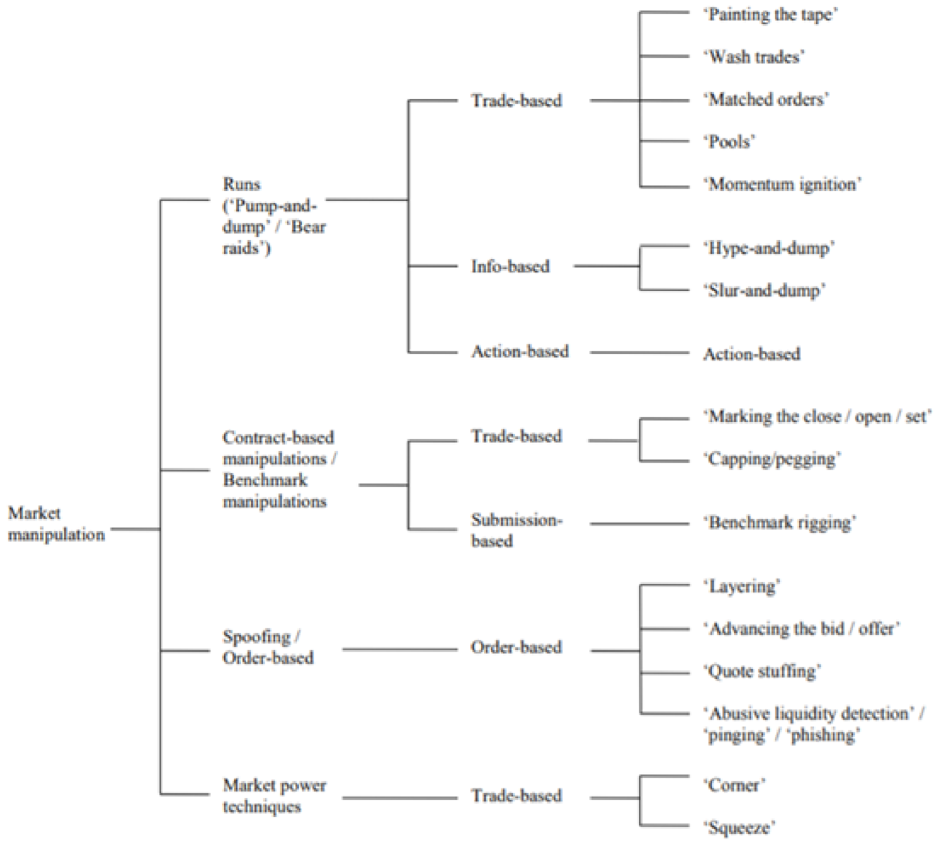


Figure 1: Taxonamy of Manipulation Techniques by Putnins, Alexander, C & Cumming, D. (2020).

In our study we are focusing on third category in the break down by Putnins which is Spoofing or Order based manipulation technique, specifically Layering and Spoofing.

Spoof trading is popular among manipulators to generate profit. A manipulator starts the tactic by placing large ask or bid offers into the market creating false sense of increase of demand for it.[34] It is not in the manipulator intention for these orders to be matched instead they will be canceled when they are about to be matched. These orders are known as passive orders and the volume of these passive sell or buy orders is usually large. The spoof orders can be implemented by either setting the passive sell price lower than the current ask price or setting the passive price higher than the current bid price.[34]

In spoofing patterns, a trader enters a single visible order, that impacts the bid offer and liquidity of the stock. Shortly before that first order is canceled, the same trader (or partner) executes a trade on the opposite side of the market. This trading behavior is manipulative because the order is matched at a better price than the trader was likely to obtain before the first order(s).[27]

Yi Cao, Yuhua Li et al, referred to specific spoofing characteristics identified by Eun Jung Leea , Kyong Shik Eomb & Kyung Suh Park as they analyzed interday order and trade data from the Korea Exchange. According to their analysis there are specific traits for spoofing manipulative behavior: the order price woule be away “6 basis points” from the current Bid/Ask offer and it’s size compared to the previous day’s average order size would be the double and would usually be canceled after more than 30 minutes. [9]

Overall spoof trading utilizes a large volume and a passive quote to cause an impact. This can be showed graphically as follows where a three-level order book is started with an offer at the best bid, pb1 and best ask, pa1 and the dotted lines represent canceled orders.[1]

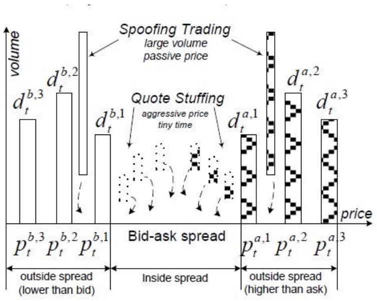


Figure 2: Example illustrating Spoofing & Quote stuffing by Zhai, J., Zhai, J., Cao, Y., Cao, Y., Ding, X., & Ding, X. (2018).

Layering is another form of spoofing where the trader enters several orders on one side of the market at multiple price ranges, to impact the spread average away from those multiple orders. After that the same trader (or a partner) executes a trade on the opposite side of the market[27]. Plot below demonstrates an example. [27]

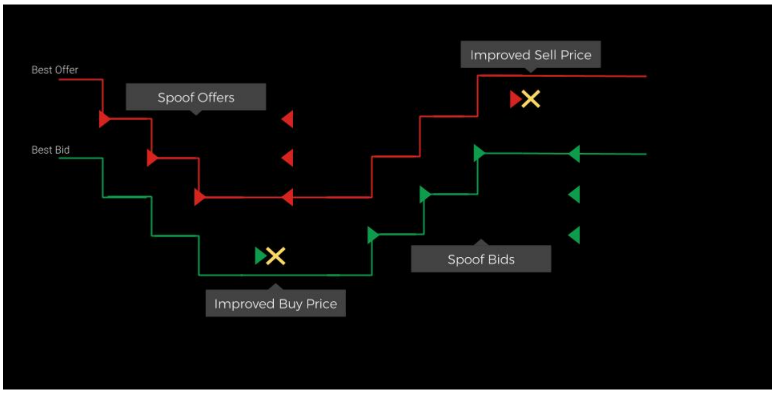


Figure 3: Layering Example from “What is the difference between layering and spoofing? (2017, June 12)”

In contrary to other regulators FIRNA use “layering” to describe entering multiple non-bona fide orders at multiple price levels while they use “spoofing” to describe entering one or more non bona fide orders at the top of the order book, thus creating a false demand for the stock. Other regulators use the two terms interchangeably. [27]

Part II
Literature Review

3 Stock Price Manipulation Detection

Stock Price manipulation detection in the literature can be divided into three stages; type of input data; feature engineering or feature extraction and detection model. Since we are interested in studying disruptive trading behaviors, it is expected to see Bid/Ask offers and volumes that are off normal trend to impact the direction of the NBBO. Such offers tend to produce a sawtooth, square wave or pulses as it has been seen in real manipulation cases.[4] Accordingly, the problem of price manipulation detection can be turned into a problem of anomaly detection in Bid/Ask price and volume time series. Since the manipulative offers usually occur in very small time intervals, the focus of our problem is to detect anomalous offers in intraday bid/ask price and volume time series.[4] There is no specific time scale to our data as it is discretely measured in terms of bid/ask price and volume offers (tick data). We can have different numbers of tick data per minute for each stock. The price fluctuations triggered by the manipulation strategies are the unusual short-termed oscillations with different amplitudes around the equilibrium level of the price. These oscillations and change in amplitude is what we try to isolate using signal analysis for feature extraction. [4]

3.1 Input Variables

Previous studies used either trade-based data such as price, volume & Bid/Ask spread as input for their models or characteristic firm specific features such as market capitalization and betas while others used market features such as index data and news data. Some authors also combine these data sources together trying to find the link between them and to enhance the results of their detection model.

Diaz et al,[6] used a mixture of all three types, where they deployed an open box or white box approach in data mining. The case study was based on published manipulation cases by the SEC in 2003, yet the data sources were different. The authors used firm specific variables such as “trading venues, market capitalization & betas”, then intraday trading information such as “price and volume within a year” and finally “news data and filing relations”[6]. Their dataset also contained market data as benchmarking for non manipulated data (Dow Jones Industrial Average)[6].

Golmohammadi et al[10] built on this study using the same dataset yet instead of using the stock prices directly as a feature in their modeling they decided to use the percentage change of price (i.e. return) or Log return. Authors argue that even though price is on of the most important variables that should be under surveillance to detect market manipulation, it should not be used in its raw form. Since that according to them the size of the company nor the revenue is reflected in the price of a stock. [10]

Our study is based on trading data, bid/ask prices and volumes to focus on disruptive trading behavior. Yet this data on its own cannot purely say if this is manipulation or not. Deeper investigation needs to be done to prove the intention of the investor and to find connections between different investors implementing the scheme if any.

3.2 Feature Engineering/Dimensionality Reduction

3.2.1 What is a feature?

According to Amanda Casari & Alice Zheng "A feature is a numeric representation of raw data"[22]. Hence there are many ways to transform raw data into numeric representation, which then relates to why features must derive from the type of data that is available. Also features can impact the model selection as some models are more appropriate for some types of features and the opposite. Generally, features can be categorized as: "relevant, irrelevant, or redundant"[3]. The number and dimensionality of features also has an impact, if the features are not informative enough (too little features) or if there are too many features (irrelevant ones) the model will not be able to generate the desirable outcome.[22]

Accordingly, authors define feature engineering as "the process of formulating the most appropriate features given the data, the model, and the task" [22]Based on the Machine learning work flow created by Amanda Casari & Alice Zheng, features and models sit between raw data and the desired insights (see Figure 5). According to the work flow we don't only pick the model, but we also choose the features, and both choices impact each other. Choosing good features make the subsequent modeling step easier and the model more accurate. Bad or non-related features can impact the model accuracy and outcome.

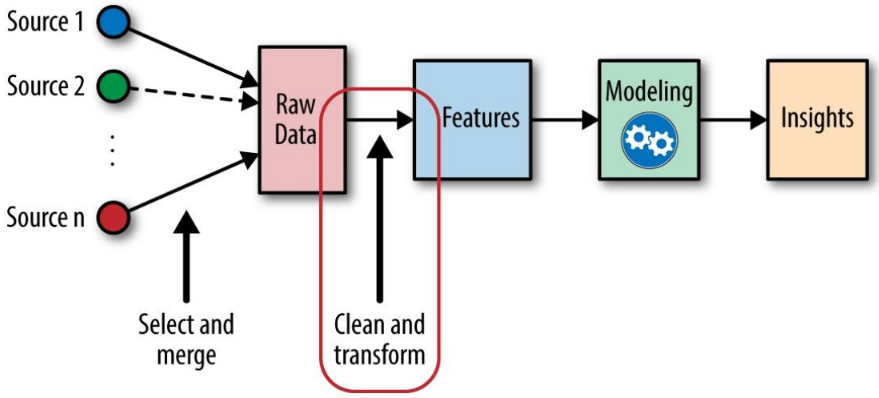


Figure 4: Feature Engineering in ML Workflow by Zheng, A., Casari, A., & Safari, an O{2019}[21]

3.2.2 What is Feature engineering?

Sinan Ozdemir; Divya Susarla defined Feature engineering as "the process of transforming data into features that better represent the underlying problem, resulting in improved machine learning performance"[21] . Authors defined 3 main steps in feature engineering, the process of transforming data which does not have to be applied to raw data only but also can be applied to preprocessed data as well. Second step would be focusing on importance of “better representing the underlying problem” . As we apply these techniques, we should not lose sight to the bigger picture and the link between these features and the problem under consideration. Then moving to resulting improvements in Machine learning as the authors highlight how the eventual goal of feature engineering is to obtain data that our learning algorithms will be able to extract patterns from and use.

Also In machine learning, dimensionality refers to the number of features (i.e. input variables) in our dataset. As mentioned earlier when the number of features is very large relative to the number of observations, certain algorithms struggle to train effective models. This is called the “Curse of Dimensionality,”.[35]

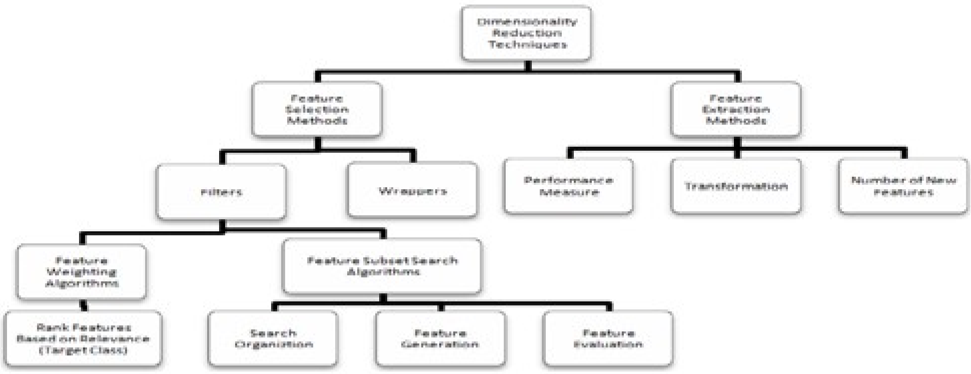

Dimensionality Reduction Techniques
Feature Selection Methods
Feature Extraction Methods
Filters
Wrappers
Performance Measure
Transformation
Number of New Features
Feature Weighting Algorithms
Feature Subset Search Algorithms
Rank Features Based on Relevance (Target Class)
Search Organization
Feature Generation
Feature Evaluation

Figure 5: Feature Engineering break down in literature by Khalid, S., Khalil, T., & Nasreen, S. (2014). [3]

Feature selection and Feature extractions are part of Feature engineering or Dimensionality reduction. In Feature Selection; one selects only those input dimensions that contain the relevant information to answer our particular question or problem while feature extraction is a more general method in which one tries to transform the space of input data to a lower dimensional subspace that keeps most of the relevant information [3]. Feature extraction and selection methods are either used in combination or separately to improve performance of the model. [3]

Features extraction can be used to reduce complexity and give a simple representation of data representing each variable in feature space as a linear combination of original input variable. An example of a widely used feature extraction approach is Principle Component Analysis (PCA) introduced by Karl. [3]

We will be combining two feature engineering methods to both extract important features from our raw data and reduce the dimensionality of the extracted features to enhance the outcome from our ML models.

3.2.3 Why treat Financial data as Signal data?

It is very common to employ different scientific concepts in different domains if the concept is valid and related, this applies to signal processing on financial data. A signal is a "physical quantity that changes with time, distance, or any independent variable". The typeof the signal depends on thetool used to extract it from the object under study. [19]

The signal originally detected is considered the raw data which can further be processed to extract more details and information. A raw signal in the financial market can be the price of the security, or its volume traded in a particular time frame. A raw signal or processed signal is a dependent variable that changes based on the changes of the independent variable, which is usually taken to be the time in the financial market. [19]

Accordingly we will be discussing signal processing application on financial data as a method of feature extraction.

3.2.4 Signal Processing:

In 1807, a French mathematician Joseph Fourier showed that any practical signal can be expressed as the sum of several sine waves, this summation is called after him as Fourier series. This transform uses sine and cosine as its bases to map a time domain function into frequency domain. [19]

Using this hypothesis, we can rewrite any financial data as a sum of sine waves. Fourier Analysis makes us choose between time or frequency. But usually, we would rather like to accurately know both time and frequency. Thus the Short-time Fourier transform and wavelet analysis are proposed. [19]

Financial data is an example for non-stationary time series, meaning that the statistical properties of the data change over time [36]. These changes are caused by different longer term business and economic cycles and in the short term by demand-supply microstructures [37]. Accordingly feature extraction method should be applicable to the non-stationary nature of the data to acquire valid input data for our detection models.

Wavelet Analysis Non-stationary time series analysis has gained interest in the recent decades in different applied sciences. In order to extract various components (or features) from non stationary time series data, several decomposition methods were developed.[20] Which allows for an improved interpretation of variability and changes in the data. Wavelet transform (WT) has been successfully applied over wide range of fields (Figure 6) to decompose the non-stationary TS into time-frequency domain. [20]

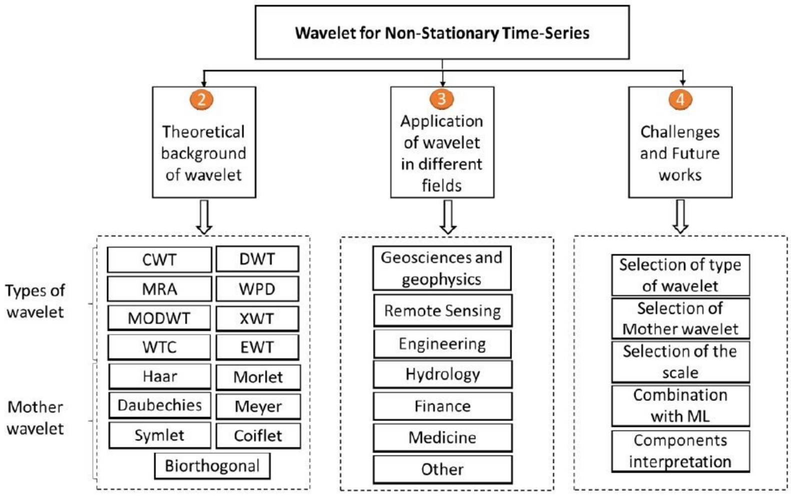

Wavelet for Non-Stationary Time-Series
2
Theoretical background of wavelet
3
Application of wavelet in different fields
4
Challenges and Future works
Types of wavelet
CWT
DWT
MRA
WPD
MODWT
XWT
WTC
EWT
Mother wavelet
Haar
Morlet
Daubechies
Meyer
Symlet
Coiflet
Biorthogonal
Geosciences and geophysics
Remote Sensing
Engineering
Hydrology
Finance
Medicine
Other
Selection of type of wavelet
Selection of Mother wavelet
Selection of the scale
Combination with ML
Components interpretation

Figure 6: Wavelet Review by Rhif, M., Ben Abbes, A., Farah, I., Martínez, B., & Sang, Y. (2019)[20]

Wavelet transform involves representing general functions in terms of simple, fixed building blocks at different scales and positions, which is named "wavelet" as first suggested by Yves Meyer and Jean MorIet [38]. These wavelets when stretched and translated, allow flexible resolution in both frequency and time. To analyze high frequency signals the window is narrowed and widened when low frequency signals need to be searched (Lau and Weng 1995).[19]

The wavelet transform (WT) was found useful for analyzing signals that are described as a “periodic, noisy, intermittent, transient and so on”. Its ability to examine the signal simultaneously in both time and frequency in a different manner from the traditional short-time Fourier transform (STFT) allowing for a wider application .[15]

Figure 7 below shows how the wavelet can be manipulated in two ways. It can be moved to various locations on the signal as shown in Figure 7.b and it can be stretched or squeezed as in figure 7.c which refers to scaling.[15]

Figure 8 shows a schematic of the wavelet transform which basically quantifies the local matching of the wavelet with the signal[15]. If the wavelet has a good match with the shape of the signal at a specific scale and location, then a large transform value is obtained. If, on the other hand the wavelet and the signal do not match well, a low transform value is obtained[15]. The transform value is then plotted on a two dimensional transform plane (bottom of figure 8). Te transformation plane is then filled up by computing the transform at various locations of the signal and for various scales of the wavelet. This is applied in a “smooth continuous way for the continuous wavelet transform (CWT) or in discrete steps for the discrete wavelet transform (DWT)”[15].

This plot builds up to be the Scalogram that we would be seeing later as we apply CWT to our data.

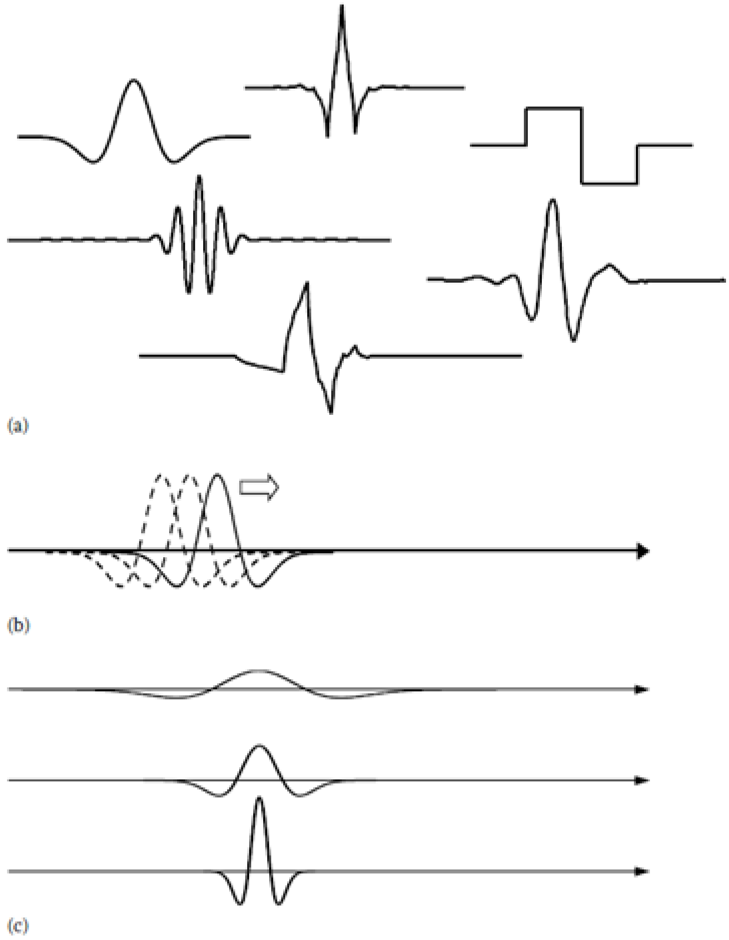


Figure 7: The Little Wave: (a) some wavelets, (b) location and (c) Scale by Addison, P. S. (2002)[15]

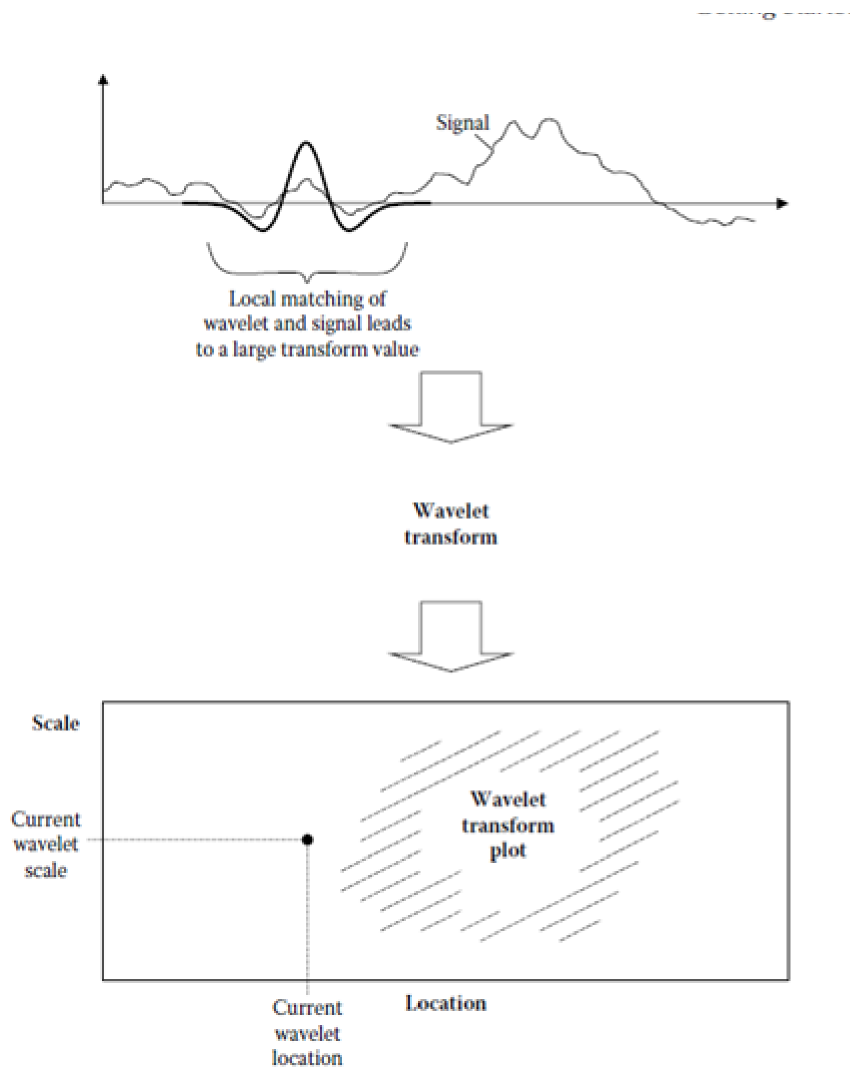


Figure 8: The Wavelet, The signal and the Transform by Addison, P. S. (2002)[15]

CWT Vs DWT Both Continuous (CWT) and Discrete (DWT) Wavelet Transform alter signals from time domain to the time-frequency domain. Yet they do it differently, the CWT uses scaling and translation to decompose a signal x(t) into a set wavelets as basis function w*(t), generated from a single basic wavelet known as mother wavelet. The obtained decomposition co-efficient, represents the original signal in some particular aspects and can be used to extract useful features. [16]

However, if scales are randomly selected or poorly selected, the resulting coefficients may reflect one original aspect of data, but other aspects might be inevitably lost[16]. Thus in our study we test the impact of using two different range of scales on the outcome of our classifiers.

The DWT on the other hand analyzes the signal at different frequency bands with different resolutions by decomposing the signal into a coarse approximation and detail information. A signal is completely decomposed into its detailed version (high frequency components) and smoothed or approximated versions (low frequency components) by employing two sets of functions, called scaling functions (low pass) and wavelet functions(high pass).

From this description, it is found that DWT mainly focuses on the information on discrete time-scale domains, while CWT extracts detailed features more efficiently by choosing the proper scale parameter.[16] Therefore, CWT is more suitable for feature extraction task in which we expect to obtain transformation features that can significantly impact the differentiation into different classes. It is also important to notice the impact of mother wavelet selection as it plays an important role as much as selecting scale parameter in signal processing or feature extraction. [16]

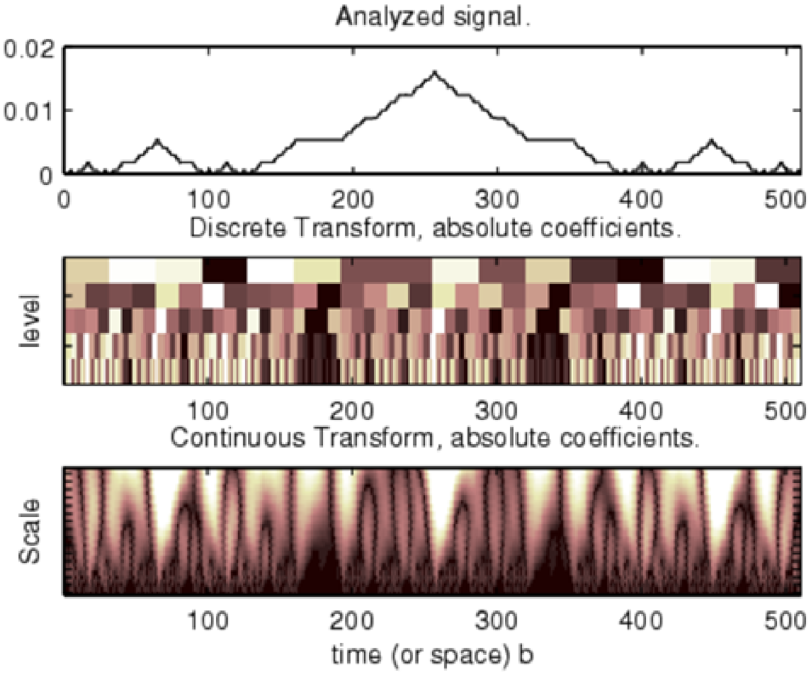


Figure 9: CWT Vs DWT from Wavelet Toolbox Matlabhelp

Application of WT in Engineering Industrial Applications Our proposed methodology was similarly applied in industrial engineering, yet with a slightly different dimensionality reduction technique . Chattopadhyay and Konar [16]discussed CWT feature extraction capabilities and experimentally verified both (CWT) and discrete wavelet transform (DWT) from the point of view of fault diagnosis of induction motors. [16]

Initial results of the redundant and high dimensionality information of CWT made it computationally in-efficient. Yet, using greedy-search feature selection technique (Greedy- CWT) the redundancy was eliminated to a great extent and found much superior to DWT technique. [16]

The feature selection technique applied by the authors has enabled determination of the most relevant CWT scales and corresponding coefficients. Thus, the inherent limitations of CWT like proper selection of scales and redundant information were avoided. .[16]Below is a comparison the authors created for their work against other research in the literature[16]. This shows the increasing interest of WT in the industrial applications.

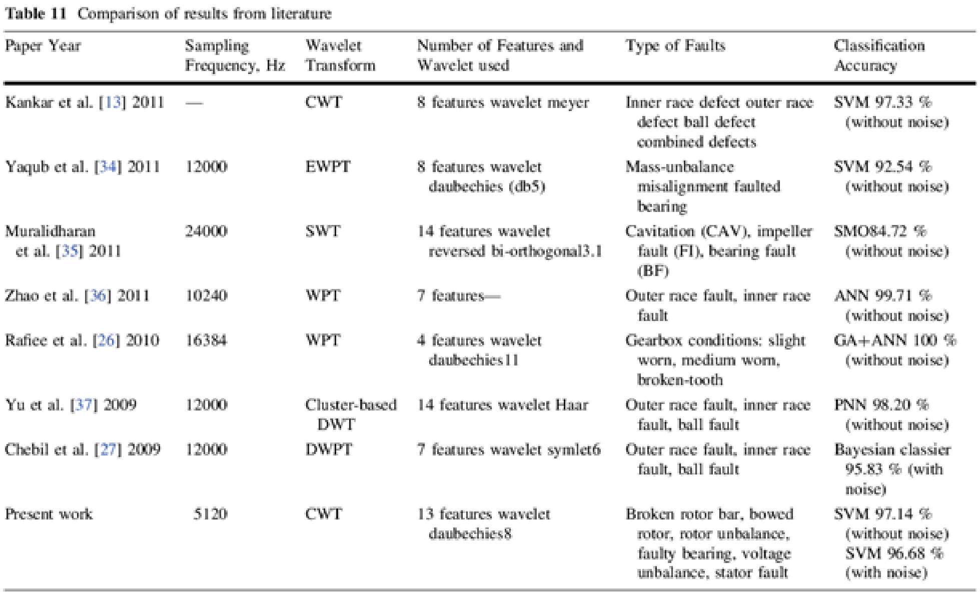


Table 11 Comparison of results from literature

Paper Year	Sampling Frequency, Hz	Wavelet Transform	Number of Features and Wavelet used	Type of Faults	Classification Accuracy
Kankar et al. [13] 2011	—	CWT	8 features wavelet meyer	Inner race defect outer race defect ball defect combined defects	SVM 97.33 % (without noise)
Yaqub et al. [34] 2011	12000	EWPT	8 features wavelet daubechies (db5)	Mass-unbalance misalignment faulted bearing	SVM 92.54 % (without noise)
Muralidharan et al. [35] 2011	24000	SWT	14 features wavelet reversed bi-orthogonal3.1	Cavitation (CAV), impeller fault (FI), bearing fault (BF)	SMO84.72 % (without noise)
Zhao et al. [36] 2011	10240	WPT	7 features—	Outer race fault, inner race fault	ANN 99.71 % (without noise)
Rafiee et al. [26] 2010	16384	WPT	4 features wavelet daubechies11	Gearbox conditions: slight worn, medium worn, broken-tooth	GA+ANN 100 % (without noise)
Yu et al. [37] 2009	12000	Cluster-based DWT	14 features wavelet Haar	Outer race fault, inner race fault, ball fault	PNN 98.20 % (without noise)
Chebil et al. [27] 2009	12000	DWPT	7 features wavelet symlet6	Outer race fault, inner race fault, ball fault	Bayesian classier 95.83 % (with noise)
Present work	5120	CWT	13 features wavelet daubechies8	Broken rotor bar, bowed rotor, rotor unbalance, faulty bearing, voltage unbalance, stator fault	SVM 97.14 % (without noise) SVM 96.68 % (with noise)

Figure 10: Comparison of results from literature for WT in Industrial applications by Chattopadhyay, P., & Konar, P. (2014)[17]

3.3 Data Mining in Price Manipulation Detection (Detection model)

In this section, we will discuss the different data mining techniques that were implemented in the literature to detect stock prices manipulation. Research will be grouped by the type of Machine learning used and the different feature extraction methodologies applied.

Implementing data mining algorithms in manipulation detection is a fairly new approach but in the past few years there has been an increase interest in it recently. Golmohammadi, Zaiane, & Diaz,[10] identified five main categories according to which data mining can be applied to stock manipulation detection:

1- Social Network Analysis: this includes detecting brokers or traders accounts that impact or manipulate the market.

2- Visualization: these tools go beyond the normal plots as they provide interactive tools to interact with the data.

3- Rule Induction produces a set of rules or limits that can be used by regulators.

4- Outlier Detection: anomaly detection or outlier detection is mainly focusing on detecting the abnormal behavior or inconsistent behavior.

5- Applying supervised learning methods for pattern recognition: the target of using this method is detecting patterns that are like previous manipulative trends.

Supervised Machine Learning

Supervised machine learning is the use of Algorithms to study patterns or externally labeled data to produce a hypothesis that can be then used to predict future samples[7]. There are two types for supervised machine learning, classification and regression. Classification is mapping new data into one of predefined classes by learning a function, while regression uses a function to map or predict a variable based on the relation between different attributes that were used to generate this function. [7]

Below we will describe some of the popular Supervised ML models used in the literature then we will go through different research to understand more about their applications and outcome in manipulation detection.

K-Nearest Neighbour (KNN) classifier assigns the input data to the class that has the most similar data points from the training data used. The standard version of KNN is non-parametric classifier where neighbors have equal vote. Accordingly the class having the maximum number of voters among the K neighbors is chosen. [13]

Conceptually, each point is plotted in a high-dimensional space, where individual variables corresponds to each axis in the space. when a new data point is introduced to the model we want to find the K nearest neighbour which means the most similar data. The Number K is important in this method and it is usually selected as the square root of N, the total number of points in the training data set.[13]

Support Vector Machine (SVM) is a kernel-based classification method based on statistical learning theory. Input data is transformed by the kernel (function) into a high dimensional space. Functions can be either linear as in dot product or non linear such as the Gaussian or the polynomial functions. [45][46]

Decision Tree (DT) uses simple rules to segment data in the form of a decision tree. it consists of one root node, a number of internal and leaf nodes and branches. The leaf nodes indicate the class to which data will be assigned to and each internal node corresponds to a feature while the branches are conjunctions of features that lead to those classifications.[12]

Logistic Regression (LR) is a widely used statistical model that uses given set of input data either continuous, discrete or a mix of both with a binary response or target. LR is different from OLR as it calculates the changes in the logarithm of odds of the response variable, instead of the changes in the dependent variable it self.[14]

Artificial Neural Network (ANN) is a computer system that imitates the biological nervous system. it consists of simple yet highly interconnected processing elements (nodes or artificial neurons). ANN is similar to the nervous system as these nodes work in parallel, learn from the process and process the information by their dynamic state responce to the external simulation. They are also able to handle fuzzy information and capable to generalize. [14]

The process by which these nodes learn varies greatly and the architecture of the network varry accordingly. Learning can be either (i) supervised, where the network is trained using input data linked to an output vector (ii) unsupervised in which there is no output data to which the input variables can be linked, accordingly the network is used to cluster the data after it is being trained on spotting similarities among the input data, or (iii) reinforced learning in which a combination of both supervised and unsupervised learning method is conducted, this method is based on a reward given to the network on the output.[14]

Research done by Aihau li, Jiede Wu & Zhhidong Liu in 2017[7] was focusing similarly to ours on trade base manipulation detection . In their paper they used both daily and high frequency tick data of 64 manipulated stock that were identified by the China Securities Regulation Commission. The authors in the paper used different supervised ML methods including: K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT), Linear Discriminant Analysis (LDA), Quadratic Discrimination Analysis (QDA), Logistic Regression (LR) &Artificial Neural Network (ANN). The authors found Supervised ML models used to give better data in classifying daily data, yet shows poor results in tick data. They explain that this due to the difficulty of labeling tick data to be used in the training data set. out of the models used they found KNN & DT to be the best which exceed 99% of all the indexes used including accuracy, senstivity, specifcity and Area Under the Curve (AUC).[7]

Similarly SVM & ANN were used by Ogut, Dogany & Aktas[11] in their attempt to detect stock market manipulation in Istanbul Stock exchange. as statistical explanatory variables the difference between the manipulated stock, index's average daily return, average daily change in trading volume and average daily volatility were used. Performance test including accuracy, sensitivity and specificity statistics were used to compare between ANN, SVM and the results of discriminant analysis and logistics regression. Based on the labeling of manipulated data and non manipulated data used, data mining techniques were found by authors to be better suited than multivariate statistical techniques to detect securities prices manipulation. As the outcome in terms of total classification accuracy and sensitivity statistics are better than those of the statistical techniques.[11]

Moving to India Stock exchange office, a comparison between discriminant analysis, ANN- GA hybrid model and SVM was made based on the classification outcome of manipulated and non manipulated stocks.[8] This study used manipulated securities identified by Securities and Echange Board if India (SEBI) during the period from 2003 to 2009. They used explanatory variables such as liquidity, volatility, price and average trading volume. Based on this analysis the results obtained by using SVM were significantly better than the results from ANN-GA and discriminant analysis model.[8]

Comparison between KNN & SVM was also done by Yi Cao, Yuhua Li et al in 2014 [9], yet this study also addressed the non stationarity nature of the financial data. In their study the authors focused on proposing a detection model based on learning and modeling the trading behavior and further identifying the manipulative actions. This is different from other detection models that focuses on unusual changes in the market features, which is most of the time due to economic cycles, public events or market moves.

Accordingly the authors identified a gap created by the lack of a model that has the ability to monitor trading behaviors directly and this is due to the lack of accurate definition of these manipulative trading behaviors. Since that these detection models are based on labeled data of what is manipulative and what is not, a benchmark input data is needed to train them, Due to the lack of this data the authors followed the identified numerical definitions of “6 bps”, “2 times” and “30 minutes” cancellation time to synthesize a dataset as they inject manipulative data in a non manipulated dataset to create a balanced labeled one. [9]

The authors addressed as well the non stationarity of financial data by using the differencing step and the log-return to transform the data into a stationary form while maintaining its features. After transforming the data, it was used as input for two Machine Learning models KNN & OCSVM . Receiver Operating Characteristics (ROC) was used to evaluate the models performance, which is calculated from the confusion matrix. OCSVM showed better outcome across all four datasets than KNN. The authors relate this higher performance to the better description of the clusters of normal cases through better description of the boundary by support vectors.[9]

Jia Zhai a et al, proposed a hybrid-model that not only uses OCSVM but also integrates a Hidden Marcov Mode (HMM).[5] This hybrid model is used to detect disruptive trading behavior focusing mainly on detecting spoofing and Quote stuffing price manipulation tactics. The authors are focusing on live manipulation detection as the Limit order data flow is extracted and calculated as a feature to the model. OCSVM focuses on identifying abnormal trends of every single trading order “ Single Order Detection”. While the other module is “Order Sequence Detection” which addresses the problem by using extended hidden Markov model to explain the contextual relationship between sequential trading orders. HMM accordingly identifies whether these sequential changes are manipulative activities (or not). Features used are trade based data that include price, size & time from the order book. Data was synthesized by reproducing the published cases and then insert them into the data set of corresponding stocks, real tick data from NASDAQ for Apple, Microsoft, Intel and Google. Based on the evaluation of the results the authors concluded that the proposed hybrid model exhibits performance that is consistently better then that of the K-NN, GMM & LR models.[5]

Hidden Marcov Model (HMM) was also used by Coa & McGinnity in 2014.[4] Yet it was still different as the authors developed an Adaptive Hidden Marcov Model with Anomaly states (AHMMAS). Based on reliable features extracted from the patterns of the manipulated bid/ask data, the AHMMAS was proposed for detecting the anomalies in them. The AHMMAS considers the anomaly states based on the thresholds of four extracted features set by the pdfs of the features. The proposed model was compared to other benchmarked methodologies such as OCSVM, KNN & GMM and it performs better in terms of area under the ROC curve and the F measure.[4]

Diaz et al, focused on detecting intraday price manipulation by deploying an open-box or a white box approach.[6] Open Box approach refers to the explainability of the model by experts. meaning that the model input and output can be interpreted and understood by experts[39]. Authors as mentioned earlier used different data types and sources based manipulation cases published by the SEC in 2003. Their data set included over 100 million trades and 170 thousands quotes. Their workflow was split into three stages, first they used clustering algorithms to cluster manipulated data points (to label hours of manipulation data since it is not provided directly by the SEC). Second step was using and testing Decision tree classification methods, testing was done using both methods bootstrapping and jack-knife. Finally the models were sorted based on their accuracy and sensitivity. The highest classification accuracy achieves was 93%. Set of rules or flags were then generated by the authors that can be used to detect manipulation by securities investigators.[6]

As mentioned earlier, Golmohammadi et al[10] built on this study using the same dataset yet they replaced the price feature with log return. They extended the previous studies that used supervised machine learning algorithms to classify manipulated data by using this dataset. they adopted different decision tree algorithms, Naive Bayes, Neural Network, SVM and KNN. Their study defined the classification problem as predicting which class Y belongs to in$\{0, 1\}$ based on a feature set of X1, X2 etc. The data set is then divided into training (27,025 out 175,738) and testing sets. All the algorithms used by the authors outperform the benchmark significantly expect of SVM which fails achieve higher than the baseline. According to the authors it is possible by using an optimization method for the parameters to reach better results yet it induces the risk of over fitting. The Naive Bayes performs better than other algorithms with sensitivity and specificity of 89% and 83% respectively. [10]

Finally research conducted by Zhai, Cao Yi & Xuemei Ding[1] is one of the closer approach to ours. The authors proposed two models one static and one dynamic to detect various aspects of price manipulation. The models were built focusing on developers trading behaviors to detect the price manipulation. Static Model input data was presented as as multi-dimentional non stationary data. In order to model it the authors proposed a transformational method inspired from differencing step and log return approach which converts the original order data into a new measure where it shows pseudo-stationary feature. [1]Based on this two popular ML methods were used as by previous others, OCSVM & KNN to detect patterns of manipulative behavior within a large data set. On the other hand the dynamic model proposed by the authors is built to try to explain the contextual relationships between sequential trading orders. To identify such manipulation cases a model is needed that analysis sequential relationships between aggressive orders and the associated time series information. Accordingly Authors contribution was highlighted in the proposal of wavelet transformation based (DWT) feature extraction approach which extracts the short term fluctuations feature of quote stuffing tactic and the proposal of Hidden Markov model based ADM which detects manipulation as changes/oscillation in bid/ask prices as opposed to a single value.[1]

In our study we will be applying CWT and combining it with PCA before running the data to two of the known classifiers above (KNN &S VM) and also introducing XGBoost a fairly recent decision tree based classifier.

Part III

Research Methodology

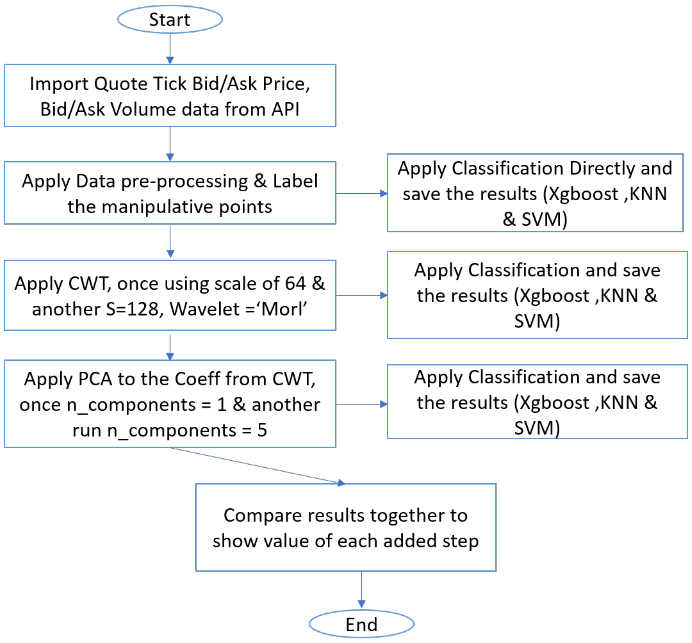
4 Detection System:

The detection system proposed in this paper is similar to the datamining work flow as it is composed of a feature extraction module from which the features are then used to train the detection model. The features are extracted from the stock bid/ask price and volume.

4.1 Proposed Work Flow

Accordingly we propose the combination of Continuous wavelet transform to transform the data into a different space then apply PCA to only use the most important features in our classification model. Thus making sure we are keeping the important features while reducing the data size and computational time.

Figure 11: Flow Chart Shows Cases that will be tested


Start
Import Quote Tick Bid/Ask Price, Bid/Ask Volume data from API
Apply Data pre-processing & Label the manipulative points
Apply Classification Directly and save the results (Xgboost ,KNN & SVM)
Apply CWT, once using scale of 64 & another S=128, Wavelet ='Morl'
Apply Classification and save the results (Xgboost ,KNN & SVM)
Apply PCA to the Coeff from CWT, once n_components = 1 & another run n_components = 5
Apply Classification and save the results (Xgboost ,KNN & SVM)
Compare results together to show value of each added step
End

4.2 Data:

4.2.1 Data Source

Data set in our research consists similarly to other authors from official cases published by the SEC from 2012 forward. Cases published by the SEC included some of the stocks that were manipulated with exact Bid and Ask numbers and timings, yet if a case was proved on 3000 stock they mention around 4-5 stocks only. Stocks were identified from two filed cases one in 2017 & the second in

2019. Also a known manipulation case that took place in Apple stock in 2013 it was published by Nanex [26]and used by Zhai, Cao Yi & Xuemei Ding[1]as a control group in their research.

High Frequency tick data is pulled from an online API with monthly subscription called Polygon.io , data is pulled through python and saved as CSV for future processing. Data extracted for same day of the actual manipulation date .Labeling for manipulated trades were done based on details mentioned in the SEC documents. The below table shows the cases published and the actual manipulation date.

Table 1: List of cases included in the dataset

Company Name	Ticker	Exchange Market	Actual Manipulation Date (MM/DD/YYYY)
Institutional Financial Markets Inc.	IFMI	NYSE	6/16/2014
Ameriserv Financial Capital Trust	ASRVP	NASDAQ	9/26/2014
Aethlon Medical Inc.	AEMD	NASDAQ	9/17/2015
Global X Gold Explorers ETF	GLDX	NYSE	1/4/2016
CHS Inc.	CHSP	NASDAQ	1/21/2016
Dawson Geophysical Co.	DWSN	NASDAQ	1/27/2017
SPAR Group	SGRP	NASDAQ	2/14/2017
Helen of Troy Limited	HELE	NASDAQ	4/1/2018
Ultragenyx Pharmaceutical Inc.	RARE	NASDAQ	5/1/2018
Helen of Troy Limited	HELE	NASDAQ	5/30/2018
Chicken Soup for the soul	CSSE	NASDAQ	7/16/2018
Chicken Soup for the soul	CSSE	NASDAQ	7/17/2018
Craft Brew Alliance	BREW	NASDAQ	9/27/2018
Cabela's Incorporated	CAP	NYSE	7/20/2015
International Business Machines Corporation	IBM	NYSE	5/20/2013
Cerner Corporation	CERN	NASDAQ	11/1/2012
Apple	AAPL	NASDAQ	7/10/2013

4.2.2 Data Description

Given that our dataset is constructed of different 15 stocks, each stock has its own Bid & Ask prices which gives a wide range of prices and creates inconsistency in the data. For example if a stock NBBO is around 2$ while another stock is around 400$ (as shown in the figure below) this will impact the anomaly detection process. We can see in Figure 12 how Apple stock has the highest NBBO while other stocks vary.

Accordingly all bid/ask prices were normalized to have consistency in the data and still be able to identify manipulative trades. Our full data set consists of 4 columns (Bid/Bid Size, Ask/Ask Size) of tick data adding up to around 278.62 K data points per column . These cases were focusing on layering and spoofing manipulation trades with total manipulated bid/bid size quotes of 490 and total of 419 manipulated ask/ask size trades.

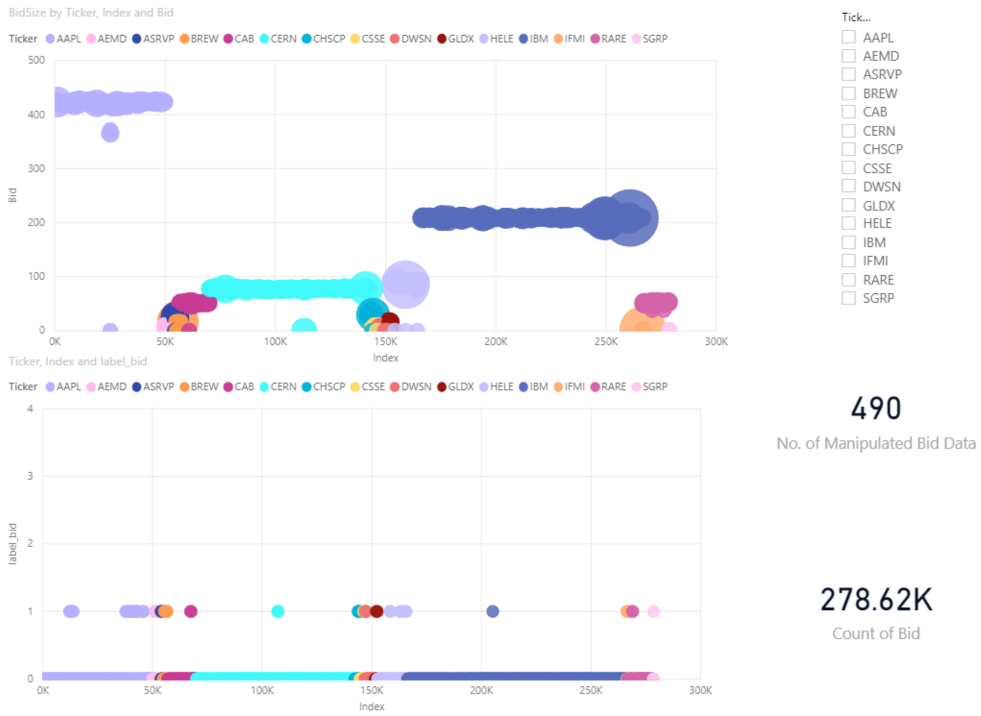
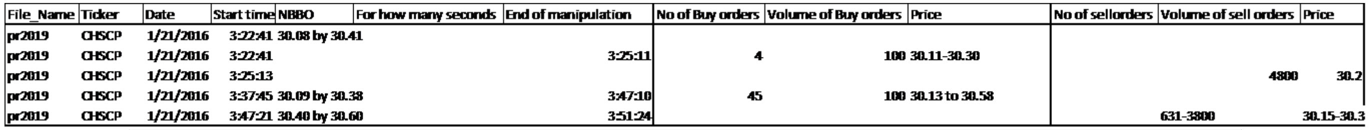


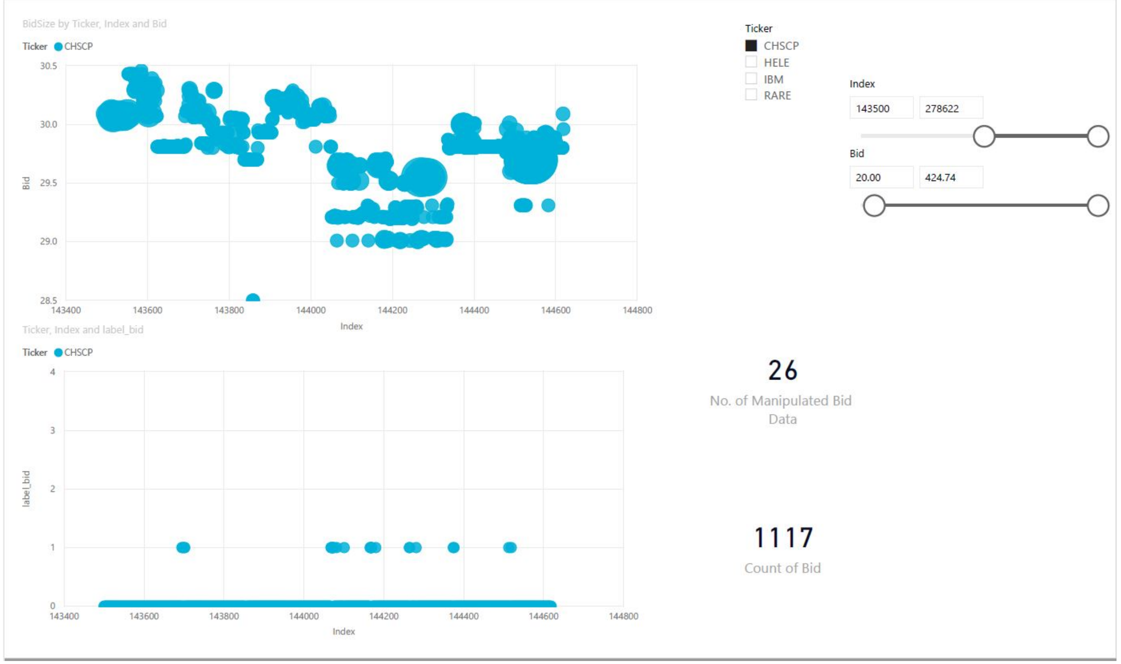
Figure 12: Figure above shows Bid price offer and the size of the marker shows the bid size, The second plot shows the labeled data (0,1)

Manipulation example from our data:

From what was found in analyzing our dataset. manipulators tend to put a limit order data then start sending non bona fid orders to impact the price in the direction that would profit them and match their limit order data. Below is an example of manipulation that took place in 2016 for CHS Inc. We can see the impact in the NBBO after the disruptive trading behavior taken by manipulators.

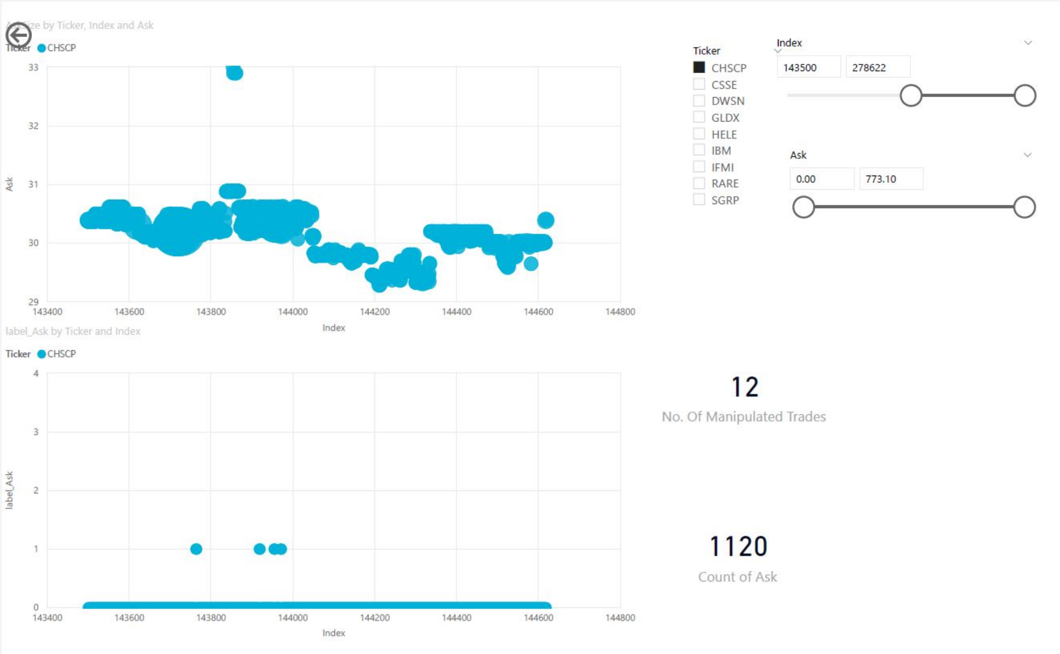


File_Name	Ticker	Date	Start time	NBBO	For how many seconds	End of manipulation	No of Buy orders	Volume of Buy orders	Price	No of sellorders	Volume of sell orders	Price
pr2019	CHSCP	1/21/2016	3:22:41	30.08 by 30.41								
pr2019	CHSCP	1/21/2016	3:22:41			3:25:11	4	100	30.11-30.30			
pr2019	CHSCP	1/21/2016	3:25:13								4800	30.2
pr2019	CHSCP	1/21/2016	3:37:45	30.09 by 30.38		3:47:10	45	100	30.13 to 30.58			
pr2019	CHSCP	1/21/2016	3:47:21	30.40 by 30.60		3:51:24					631-3800	30.15-30.3



Above Plot shows the Bid data on the manipulation day, different averages can be seen. One before disruptive trading behavior, second one at lower values due to it's impact then going back to normal. The size of the dot shows how big was the volume of the Bid offer. Both has an impact on the NBBO as they create the false feeling of increasing demand. The second plot shows the trades labeled 1 (manipulated) we can see the impact of these trades as they happen on the NBBO.

Figure 13: Actual Manipulation example from our dataset



Ticker

■ CHSCP
□ CSSE
□ DWSN
□ GLDX
□ HELE
□ IBM
□ IFMI
□ RARE
□ SGRP

Index

143500 278622

Ask

0.00 773.10

12

No. Of Manipulated Trades

1120

Count of Ask

Figure 14: Actual Manipulation example from our dataset

Above Plot shows the Ask data on the manipulation day, different averages can be seen. One before disruptive trading behavior, second one at lower values due to it's impact then going back to a lower average than the previous normal The size of the dot shows how big was the volume of the Ask offer. The second plot shows the trades labeled 1 (manipulated) we can see the impact of these trades as they happen on the NBBO.

4.3 Feature Extraction

AS mentioned earlier that feature extraction helps us transform our data into features that better represent the underlying problems. Accordingly this research applies Signal processing and specifically Continuous wavelet transform for feature extraction. It was applied using Pywavelets library in python, using "Morl" wavelet and sensitivities were made on the scale range.

CWT as shown before was used by authors in different domains, from industrial engineering to medical and finance applications. The main idea behind it is to decompose the prices and quote size signals into different sub signals or frequencies. These coeff are then used as features to train and test classification models since that anomalies show as abrupt changes in these coeff.

Pywavelets library[40] in python supports 1D CWT, it requires minimum input as shown below and returns two arrays (Frequencies & Coeff).

pywt.cwt*(data, scales, wavelet)*

data : array_like Input signal (in our case it is our Bid price/ Bid Size/Ask Price & Ask Size data).

scales : array_like The wavelet scales to use. Scale is the inverse of frequency, high scales means lower frequencies. In our model we tested both 64 & 128 to compare both.

wavelet : Wavelet object or name Wavelet to use, this library supports 7 CWT wavelets out of which 3 (Mother wavelets) are tested and the Morl gave the highest results in classification.

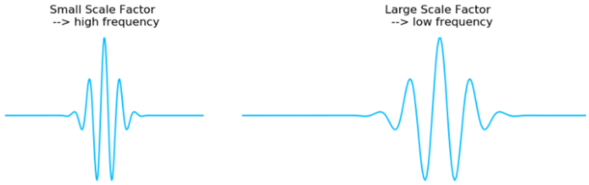


Figure 15: Demonstration of shrink and scaled Morl Wavelet, Scale is inversely proportional to frequency by Feike, S. (2020, February 10)[23]

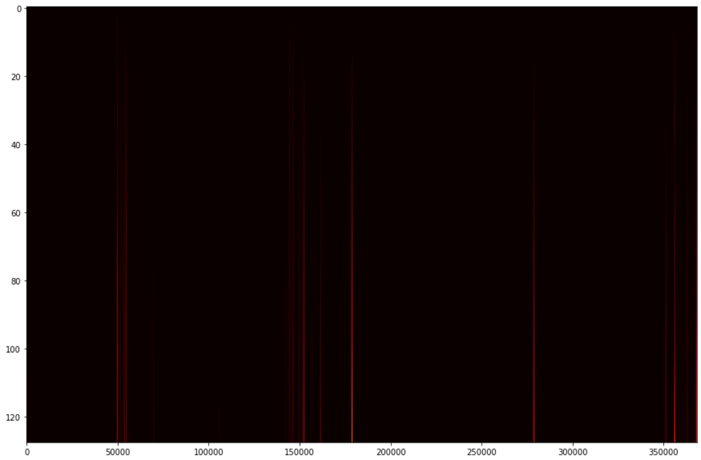


Figure 16: Scalogram Example applying scale 1-> 128 on our dataset Bid Prices

Figure 14 represents a scalogram, which helps visualize the outcome of CWT. it acts as a visualization for the coeff vs scale. Anomalies can be seen as the breaks in this scalogram, some researches use these scalograms as direct input to Neural Networks for pattern detection. In our case we are using the coefficients directly as features for classification.

4.4 Dimension Reduction (PCA)

As mentioned earlier Principle Component Analysis PCA is one of the widely used techniques in feature engineer. it is a mathematical algorithm that reduces the dimensionality of the data while maintaining most of the variation in the original data. To accomplish this reduction it identifies directions, along which the variation in the data is maximal (Principle Components). Data can still be represented by relatively few numbers by using these components. PCA is applied using Sklearn library [41]in python as shown below. it applies linear dimensionality reduction through Singluar Value Decomposition (SVD) of the data to project it to a lower space.

class sklearn.decomposition.PCA(n_components=None, *, copy=True, whiten=False, svd_solver='auto', tol=0.0, iterated_power='auto', random_state=None)

n_components: int, float or str, default=None Number of components to keep. if n_components is not set all components are kept.

Number of components in our research was once set to 1 and another time set to 5 in order to test how sensitive the results are to it, also to test how much the algorithm running time will increase.

4.5 Detection Model

The main contribution of this research is in the feature extraction segment of our workflow, yet comparing different detection (Binary classification) Machine learning models can also be beneficial.

Most of the research done previously focused more on SVM & KNN as shown earlier. In our research we are proposing to also compare these models to XGBoost a tree ML classification model.

XGBoost, is scalable machine learning system for tree boosting[42]. The system is available as an open source package. XGBoost became very popular recently as it showed great performance as shown in the challenges hosted by Kaggle, the machine learning competition site. XGBoost was used 17 times among the 29 challenge winning solutions published at Kaggle's blog in 2015.[42]

It is an algorithm that uses ensemble of decision trees where new trees fix errors of those trees that are already part of the model. Trees are added until no further improvements can be made to the model. It is a faster improvement to gradient boost trees. In Gradient boosting greedy search is used to decide which leaf or branch to open next based on the least loss function value. This is similarly done in XGboost yet it makes some regularization that makes the process much faster. It looks at feature distribution across all data points in a leaf and accordingly reduces the search space of possible feature splits. [43]

XGBoost also allows us to compensate for the imbalanced dataset we are using. It allows us to identify a weight number as an input parameter in our function. The weight number scale is called *scale_pos_weight[44]* and one of the approximation calculation for it could be the following:

scale_pos_weight = *total_negative_examples / total_positive_examples*

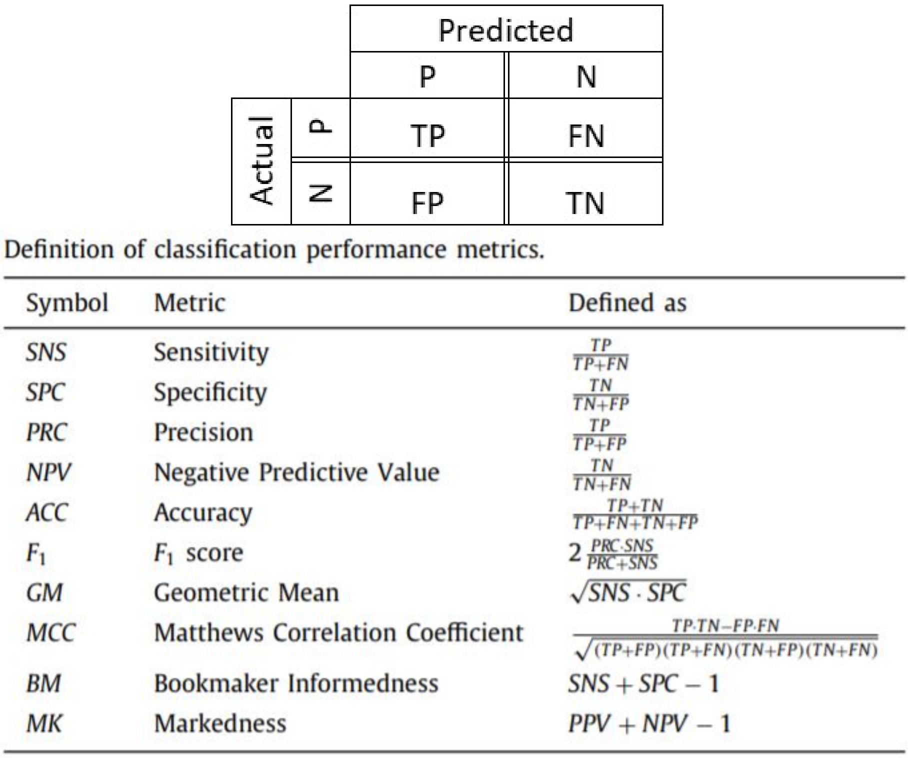
Where the negative examples for us is the 0 or the non-manipulated data points and the positive are the 1 that we are trying to predict. In our dataset we have two Scale_pos_weight as we calculate it once for the Bid data and another time for the Ask data. As mentioned earlier our dataset has ~278.62K quote data with 490 positive point in Bid data and 419 in Ask data. This gives us weighting scales of: 567 & 664 for Bid and Ask respectively.

As discussed earlier we also included KNN & SVM to be able to compare to similar research by other authors.

4.6 Evaluation Matrix

In order to evaluate the impact of the proposed methodology we monitored and compares classification models results. In order to have comparable metrics we calculated the confusion matrix[2].

Confusion Matrix serves as a simple visualization tool that can be used to view classifier results. it helps also calculating other metrics such as Accuracy, Precision and Recall which makes it easier to compare between different models outcome. Below is an example of confusion matrix and what metrics can we calculate from it.[2]



		Predicted	
		P	N
Actual	P	TP	FN
	N	FP	TN

Definition of classification performance metrics.

Symbol	Metric	Defined as
SNS	Sensitivity	$\frac{TP}{TP+FN}$
SPC	Specificity	$\frac{TN}{TN+FP}$
PRC	Precision	$\frac{TP}{TP+FP}$
NPV	Negative Predictive Value	$\frac{TN}{TN+FN}$
ACC	Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$
F_1	F_1 score	$2\,\frac{PRC \cdot SNS}{PRC+SNS}$
GM	Geometric Mean	$\sqrt{SNS \cdot SPC}$
MCC	Matthews Correlation Coefficient	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
BM	Bookmaker Informedness	$SNS + SPC - 1$
MK	Markedness	$PPV + NPV - 1$

Accuracy is % of correctly predicted labeled data either 0 or 1.
Precision is % of Correctly predicted 1 out of all 1s predicted by the model.
Recall is % of Correctly predicted 1 out of the total 1s that re in the dataset.

Figure 17: Plot on top shows Confusion Matrix &Table below shows how each Metric can be calculated. Table by Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019) [2]

Given that we are dealing with imbalanced dataset where we care more about the positive data, we are mainly interested in measuring the probability of the classifier detecting True positive and False Positive. Accordingly we will be focusing on Precision and Recall. Accuracy will also be reported yet it will mainly be high as the greater number of points lie on one side.

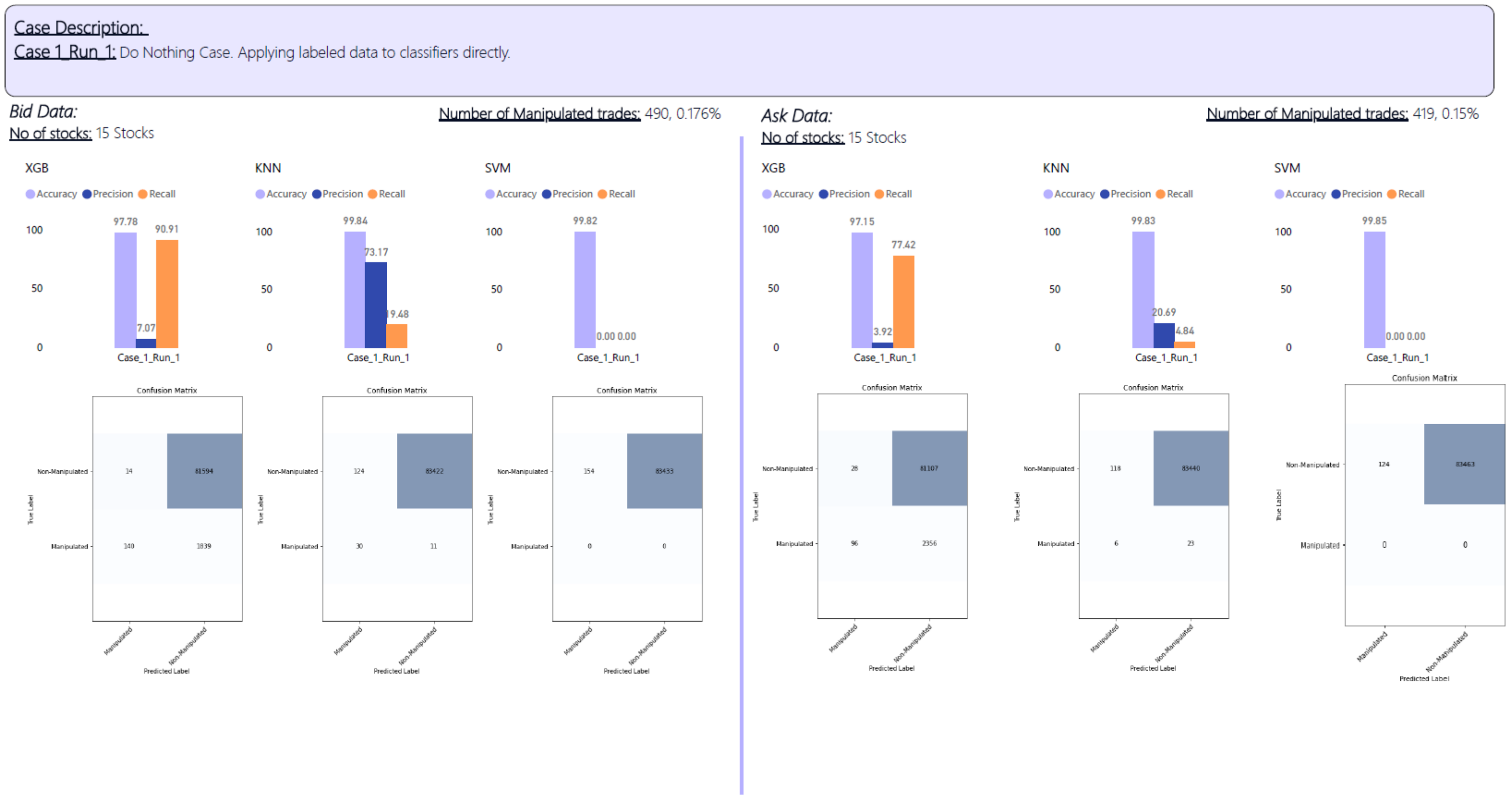
Part IV

Results & Discussion

As mentioned in our methodology section, We are comparing 3 different cases which are split into 5 test runs. We will go through each case and compare to the incremental step implemented to test the impact of each.

Our first run is our do nothing case, where we simply take our data set and apply normal statistical operation to remove nonstationarity then we apply classification directly. Some authors as mentioned earlier applied this as they were focusing on comparing different supervised ML models. In our case we will take it as the benchmark against which we will introduce feature extraction methods to enhance the outcome.

Case 1_Run_1 Results are:



Case Description:
Case 1_Run_1: Do Nothing Case. Applying labeled data to classifiers directly.

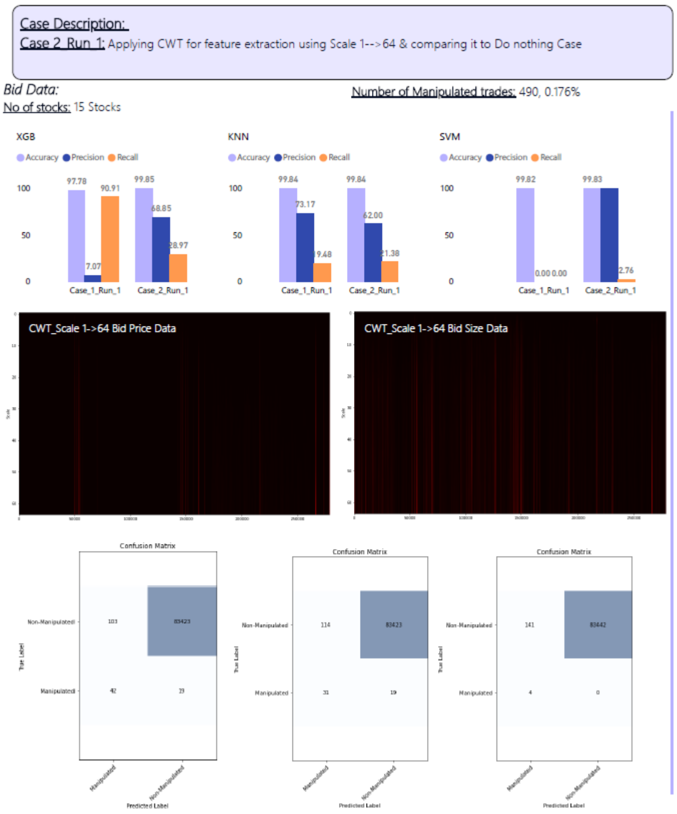
Bid Data:
No of stocks: 15 Stocks
Number of Manipulated trades: 490, 0.176%

Ask Data:
No of stocks: 15 Stocks
Number of Manipulated trades: 419, 0.15%

Figure 18: Comparison Above showing 3 classifiers results for both Bid & Ask Data

Figure 18 shows the results of running classification on our data without any feature extraction applied. It shows that KNN has the highest Precision which means that it detects correctly more positive than negative out of all my positive predictions. While XGB had the highest Recall numbers which means that it was able to predict higher % of 1s in all the manipulated data we had.

Case_2_Run_1_Bid Data



Case Description:

Case 2_Run_1: Applying CWT for feature extraction using Scale 1-->64 & comparing it to Do nothing Case

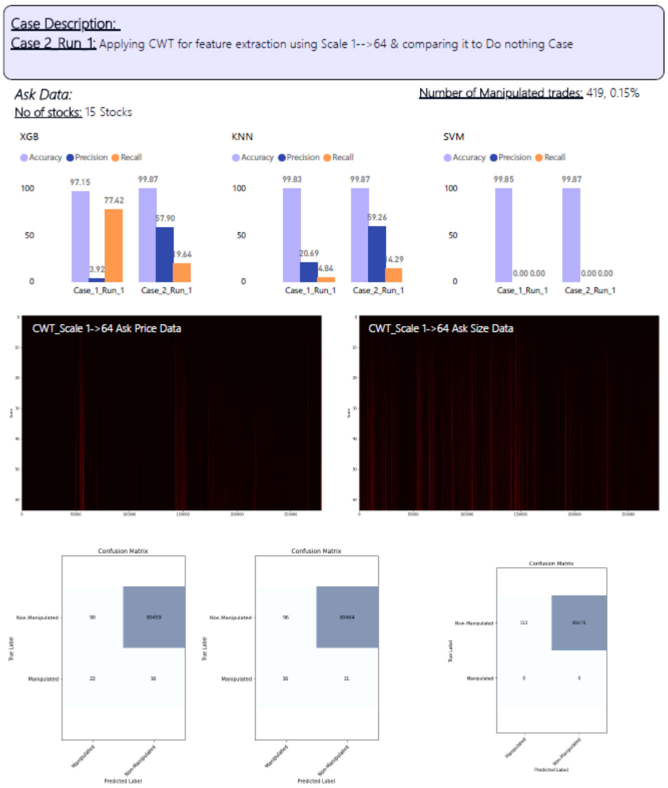
Bid Data:

No of stocks: 15 Stocks

Number of Manipulated trades: 490, 0.176%

Figure 19: Case_2_Run_1_Bid Data Classifiers Comparison

Case_2_Run_1_Ask Data



Case Description:

Case 2 Run 1: Applying CWT for feature extraction using Scale 1-->64 & comparing it to Do nothing Case

Ask Data:

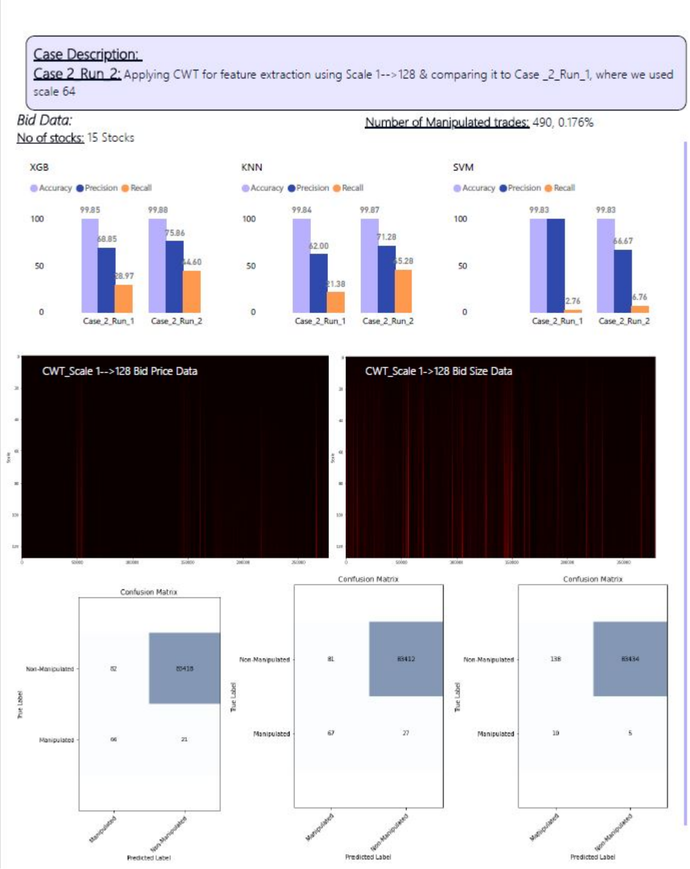
No of stocks: 15 Stocks

Number of Manipulated trades: 419, 0.15%

Figure 20: Case_2_Run_1_Ask Data Classifiers Comparison

Above plots show the impact of applying CWT of scale 64 to the data before classification for XGBoot there was a great increase in precision while a drop in recall took place. For KNN in case of bid data it was almost the same while SVM saw very good enhancement in its performance. In General results are better on Bid Data as we have more Positive incidents in the dataset allowing for better training.

Case_2_Run_2_Bid Data



Case Description:

Case 2 Run 2: Applying CWT for feature extraction using Scale 1-->128 & comparing it to Case _2_Run_1, where we used scale 64

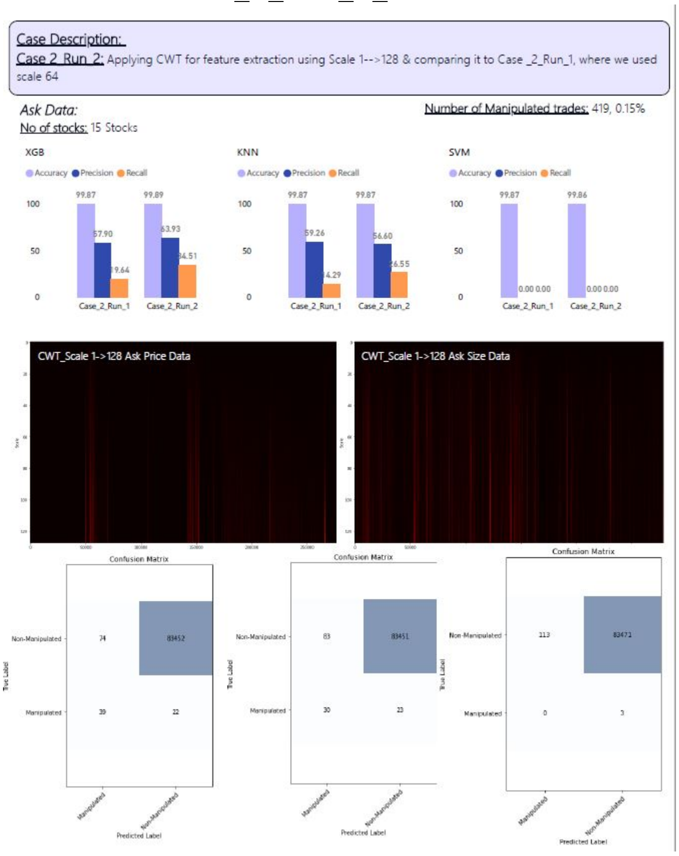
Bid Data:

No of stocks: 15 Stocks

Number of Manipulated trades: 490, 0.176%

Figure 21: Case_2_Run_2_Bid Data Classifiers Comparison

Case_2_Run_2_Ask Data



Case Description:

Case 2 Run 2: Applying CWT for feature extraction using Scale 1-->128 & comparing it to Case _2_Run_1, where we used scale 64

Ask Data:

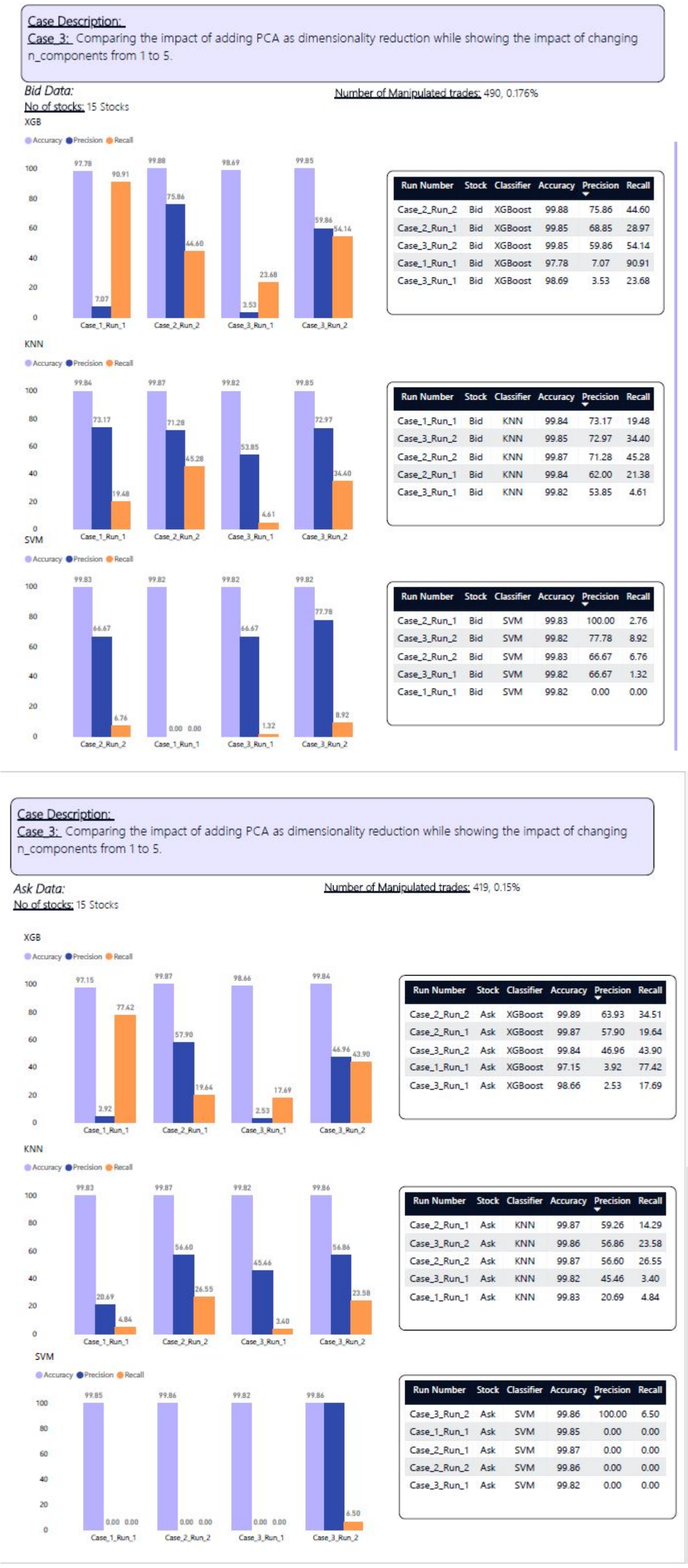
No of stocks: 15 Stocks

Number of Manipulated trades: 419, 0.15%

Figure 22: Case_2_Run_2_Ask Data Classifiers Comparison

Above comparison shows the impact of increasing scale in CWT from 64 to 128. In Case_2 second run we increased the scale in CWT to 128 which is shown on the scalogram as more change in frequencies can be seen. this resulted in a jump in performance for all three classifiers for Bid data while for Ask data XGBoost and KNN showed performance enhancement too. This shows the positive impact of CWT on feature extraction for classification and identifying abnormal trading behaviors. Yet one of the major drawbacks to this workflow is the run time, accordingly PCA will be implemented next to reduce data dimension and test if we can still maintain the good results.

Case_3



Case Description:
Case 3: Comparing the impact of adding PCA as dimensionality reduction while showing the impact of changing n_components from 1 to 5.

Bid Data:
No of stocks: 15 Stocks

Number of Manipulated trades: 490, 0.176%

XGB

Run Number	Stock	Classifier	Accuracy	Precision	Recall
Case_2_Run_2	Bid	XGBoost	99.88	75.86	44.60
Case_2_Run_1	Bid	XGBoost	99.85	68.85	28.97
Case_3_Run_2	Bid	XGBoost	99.85	59.86	54.14
Case_1_Run_1	Bid	XGBoost	97.78	7.07	90.91
Case_3_Run_1	Bid	XGBoost	98.69	3.53	23.68

KNN

Run Number	Stock	Classifier	Accuracy	Precision	Recall
Case_1_Run_1	Bid	KNN	99.84	73.17	19.48
Case_3_Run_2	Bid	KNN	99.85	72.97	34.40
Case_2_Run_2	Bid	KNN	99.87	71.28	45.28
Case_2_Run_1	Bid	KNN	99.84	62.00	21.38
Case_3_Run_1	Bid	KNN	99.82	53.85	4.61

SVM

Run Number	Stock	Classifier	Accuracy	Precision	Recall
Case_2_Run_1	Bid	SVM	99.83	100.00	2.76
Case_3_Run_2	Bid	SVM	99.82	77.78	8.92
Case_2_Run_2	Bid	SVM	99.83	66.67	6.76
Case_3_Run_1	Bid	SVM	99.82	66.67	1.32
Case_1_Run_1	Bid	SVM	99.82	0.00	0.00

Case Description:
Case 3: Comparing the impact of adding PCA as dimensionality reduction while showing the impact of changing n_components from 1 to 5.

Ask Data:
No of stocks: 15 Stocks

Number of Manipulated trades: 419, 0.15%

XGB

Run Number	Stock	Classifier	Accuracy	Precision	Recall
Case_2_Run_2	Ask	XGBoost	99.89	63.93	34.51
Case_2_Run_1	Ask	XGBoost	99.87	57.90	19.64
Case_3_Run_2	Ask	XGBoost	99.84	46.96	43.90
Case_1_Run_1	Ask	XGBoost	97.15	3.92	77.42
Case_3_Run_1	Ask	XGBoost	98.66	2.53	17.69

KNN

Run Number	Stock	Classifier	Accuracy	Precision	Recall
Case_2_Run_1	Ask	KNN	99.87	59.26	14.29
Case_3_Run_2	Ask	KNN	99.86	56.86	23.58
Case_2_Run_2	Ask	KNN	99.87	56.60	26.55
Case_3_Run_1	Ask	KNN	99.82	45.46	3.40
Case_1_Run_1	Ask	KNN	99.83	20.69	4.84

SVM

Run Number	Stock	Classifier	Accuracy	Precision	Recall
Case_3_Run_2	Ask	SVM	99.86	100.00	6.50
Case_1_Run_1	Ask	SVM	99.85	0.00	0.00
Case_2_Run_1	Ask	SVM	99.87	0.00	0.00
Case_2_Run_2	Ask	SVM	99.86	0.00	0.00
Case_3_Run_1	Ask	SVM	99.82	0.00	0.00

Figure 23: Case_3_Results both Bid & Ask Data

Above Comparison Shows the impact of adding PCA after CWT with two sensitivities of keeping top 1 component and keeping top 5 components. Case _3 also had two sensitivities as we dropped the case with CWT scale of 64 instead we built on using CWT with scale of 128 integrated with PCA. one of the variables that impacts the outcome of PCA is the number of components kept. Accordingly we ran two sensitivities to see which would give better results.

First run (Case_3_Run_1) was using number of components of 1, this reduced computational time tremendously yet it hurt the performance of the classifiers. Thus the second run was tried with number of components of 5. This as well kept a good running time, much less than without PCA and classifiers performance was close to Case_2_Run_2 which gave a better outcome from our do nothing case.

Part V
Conclusion

Data Mining and Machine learning can add a lot to the space of disruptive trading behavior detection. It can deal with high volume of data unlike the rule or the threshold technique currently held by regulators. As shown in our research applying signal processing to our financial raw data helps identifying the coefficients that best represented those trades. Specifically applying continuous wavelet transform CWT then PCA added to the performance of our classifiers, Also the testing of the famous and widely used tree model XGBoost gave us the flexibility of adapting our model to the imbalanced data set we have. In most of our cases we can see a higher performance of XGBoost or a very close performance between both KNN & XGBoost.

The proposed system as mentioned earlier adds the big advantage of being able to detect anomalous trades on live data feed. Yet this system might not be straight forward to apply due to these challenges: 1) High Frequency data comes at a price as it is expensive to acquire. 2) It needs high performance computers to be able to handle this big amount of data and run the process contiguously.

Yet if applied on bigger scale having such a system can reduce the time between highlighting a possible fraud and actually proving it. Also Machine Learning adds the benefit of having an updated model and generalized that can detect different instances of disruptive trading, instead of having thresholds that could go out of date.

For future studies there are three areas of improvements that can be applied. First for data input, we can acquire a bigger dataset and look for more cases published by the SEC, yet automating the labeling process to avoid human error would be favorable. As for the continuous wavelet transform application, more effort can be done to optimize and automate both wavelet and scale selection. As for machine learning section, more analysis can be done on the significance of features in more details.

References

[1] Zhai, J., Zhai, J., Cao, Y., Cao, Y., Ding, X., & Ding, X. (2018). Data analytic approach for manipulation detection in stock market. Review of Quantitative Finance and Accounting, 50(3), 897-932. doi:10.1007/s11156-017-0650-0

[2] Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216-231. doi:10.1016/j.patcog.2019.02.023

[3] Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. Paper presented at the 372-378. doi:10.1109/SAI.2014.6918213

[4] Cao, Y., Li, Y., Coleman, S., Belatreche, A., & McGinnity, T. M. (2015). Adaptive hidden markov model with anomaly states for price manipulation detection. IEEE Transaction on Neural Networks and Learning Systems, 26(2), 318-330. doi:10.1109/TNNLS.2014.2315042

[5] Zhai, J., Cao, Y., Yao, Y., Ding, X., & Li, Y. (2017). Computational intelligent hybrid model for detecting disruptive trading activity. Decision Support Systems, 93, 26-41. doi:10.1016/j.dss.2016.09.003

[6] Diaz, D., Theodoulidis, B., & Sampaio, P. (2011). Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices. Expert Systems with Applications, 38(10), 12757-12771. doi:10.1016/j.eswa.2011.04.066

[7] Aihua Li, Jiede Wu, Zhidong Liu (2017). Market Manipulation Detection Based on Classification Methods. Procedia Computer Science,122, 788-795,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2017.11.438.

[8] Thoppan, J. J., Punniyamoorthy, M., & Ganesh, K. (2017). Competitive models to detect stock manipulation. Communications of the IIMA, 15(2), 1-17.

[9] Cao, Y., Li, Y., Coleman, S., Belatreche, A., & McGinnity, T. M. (2014). Detecting price manipulation in the financial market. Paper presented at the 77-84. doi:10.1109/CIFEr.2014.6924057

[10] Golmohammadi, K., Zaiane, O. R., & Diaz, D. (2014). Detecting stock market manipulation using supervised learning algorithms. Paper presented at the 435-441. doi:10.1109/DSAA.2014.7058109

[11] Öğüt, H., Mete Doğanay, M., & Aktaş, R. (2009). Detecting stock-price manipulation in an emerging market: The case of turkey. Expert Systems with Applications, 36(9), 11944-11949. doi:10.1016/j.eswa.2009.03.065

[12] Cheng-Jin Du, Da-Wen Sun. (2008).4 - Object Classification Methods,Computer Vision Technology for Food Quality Evaluation. In Food Science and Technology,Academic Press,81-107,ISBN 9780123736420,https://doi.org/10.1016/B978-012373642-0.50007-7.

[13] Nadkarni, P. (2016). Clinical research computing: A practitioner's handbook. Amsterdam, Netherlands: Academic Press.

[14] Hajmeer, M., & Basheer, I. (2003). Comparison of logistic regression and neural network-based classifiers for bacterial growth. Food Microbiology, 20(1), 43-55. doi:10.1016/S0740-0020(02)00104-1

[15] Addison, P. S. (2002). The illustrated wavelet transform handbook: Introductory theory and applications in science, engineering, medicine, and finance. Philadelphia;Bristol, UK;: Taylor & Francis.

[16] Chattopadhyay, P., & Konar, P. (2014). Feature extraction using wavelet transform for multi-class fault detection of induction motor. Journal of The Institution of Engineers (India): Series B, 95(1), 73-81.

[17] Alexander, C. (., & Cumming, D. (2020). Corruption and fraud in financial markets: Malpractice, misconduct and manipulation. Chichester, West Sussex, United Kingdom: Wiley.

[18] Matlap help : http://www.ece.northwestern.edu/local-apps/matlabhelp/toolbox/wavelet/cwt.html

[19] K, M. D. (2003). Science of financial market trading, the. Singapore: World Scientific Publishing Company. doi:10.1142/5178#t=toc

[20] Rhif, M., Ben Abbes, A., Farah, I., Martínez, B., & Sang, Y. (2019). Wavelet transform application for/in non-stationary time-series analysis: A review. Applied Sciences, 9(7), 1345. doi:10.3390/app9071345

[21] Ozdemir, S., Susarla, D., & Safari, an O{u2019}Reilly Media Company. (2018). Feature engineering made easy (1st ed.) Packt Publishing.

[22] Zheng, A., Casari, A., & Safari, an O{u2019}Reilly Media Company. (2018). Feature engineering for machine learning (1st ed.) O'Reilly Media, Inc.

[23] Feike, S. (2020, February 10). Multiple Time Series Classification by Using Continuous Wavelet Transformation. Medium. https://towardsdatascience.com/multiple-time-series-classification-by-using-continuous-wavelet-transformation-d29df97c0442

[24] Cheboli, D. Anomaly Detection of Time Series. Facility Of The Graduate School Of The University Of Minnesota.—2010.—75 c.—[Электронный ресурс]—Режим доступа. URL: http://conservancy. umn. edu/bitstream/handle/11299/92985.

[25] Koch v. SEC, No. 14-1134 (D.C. Cir. 2015). (2015, July 14). Justia Law. https://law.justia.com/cases/federal/appellate-courts/cadc/14-1134/14-1134-2015-07-14.html

[26] Nanex (2013) Incredible, blatant manipulation in Apple stock. Retrieved from http://www.nanex.net/aqck2/ 4352.html

[27] What is the difference between layering and spoofing? (2017, June 12). Trillium Management, LLC. https://www.trlm.com/knowledgebase/makes-spoofing-different-layering/

[28] Cartea, Álvaro and Jaimungal, Sebastian and Wang, Yixuan, Spoofing and Price Manipulation in Order Driven Markets (August 2, 2019). Available at SSRN: https://ssrn.com/abstract=3431139 or http://dx.doi.org/10.2139/ssrn.3431139

[29] Orlando Cosme Jr., Regulating High-Frequency Trading: The Case for Individual Criminal Liability, 109 J. Crim. L. & Criminology 365 (2019). https://scholarlycommons.law.northwestern.edu/jclc/vol109/iss2/5

[30] COSME, O. (2019). REGULATING HIGH-FREQUENCY TRADING: THE CASE FOR INDIVIDUAL CRIMINAL LIABILITY. The Journal of Criminal Law and Criminology (1973-), 109(2), 365-394. doi:10.2307/48572780

[31] Mermin, A., & Pickard, R. F. (1947). Regulation of Stock Market Manipulation. Yale Law Journal, 56, 509-33.

[32] Markham, J. W. (2013). Law enforcement and the history of financial market manipulation. ME Sharpe.

[33] E. JACKSON, H., & E. TAHYAR, M. (2020). Fintech Law: The Case Studies. Harvard University. https://projects.iq.harvard.edu/files/fintechlaw/files/fintech_law_the_case_studies.pdf

[34] Leangarun, T., Tangamchit, P., & Thajchayapong, S. (2016). Stock price manipulation detection based on mathematical models. International Journal of Trade, Economics and Finance, 7(3), 81-88.

[35] E. (2020, June 9). Dimensionality Reduction Algorithms: Strengths and Weaknesses. EliteDataScience. https://elitedatascience.com/dimensionality-reduction-algorithms#:%7E:text=In%20machine%20learning%2C%20%E2%80%9Cdimensionality%E2%80%9D,st

[36] Ghazali, R., Jaafar Hussain, A., Mohd Nawi, N., & Mohamad, B. (2009). Non-stationary and stationary prediction of financial time series using dynamic ridge polynomial neural network. Neurocomputing (Amsterdam), 72(10), 2359-2367. doi:10.1016/j.neucom.2008.12.005

[37] Allen, F., & Gorton, G. (1991). Stock price manipulation, market microstructure and asymmetric information (No. w3862). National Bureau of Economic Research.

[38] Motard, R. L., & Joseph, B. (Eds.). (2013). Wavelet applications in chemical engineering (Vol. 272). Springer Science & Business Media.

[39] Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access, 7, 154096-154113.

[40] Gregory R. Lee, Ralf Gommers, Filip Wasilewski, Kai Wohlfahrt, Aaron O'Leary (2019). PyWavelets: A Python package for wavelet analysis. Journal of Open Source Software, 4(36), 1237, https://doi.org/10.21105/joss.01237.

[41] Pedregosa, F et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825--2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[42] Chen, T., & Guestrin, C. XGBoost: A scalable tree boosting system. arXiv 2016. arXiv preprint arXiv:1603.02754.

[43] Tseng, G. (2018, November 29). Gradient Boosting and XGBoost - Gabriel Tseng. Medium. https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5

[44] Brownlee, J. (2020, August 21). How to Configure XGBoost for Imbalanced Classification. Machine Learning Mastery. https://machinelearningmastery.com/xgboost-for-imbalanced-classification/

[45] Srikanta Mishra, Akhil Datta-Gupta (2018).Chapter 8 - Data-Driven Modeling,Applied Statistical Modeling and Data Analytics,Elsevier,195-224,9780128032794 , https://doi.org/10.1016/B978-0-12-803279-4.00008-0.

[46] Pourghasemi, H. R., & Gokceoglu, C. (2019). Spatial Modeling in GIS and R for Earth and Environmental Sciences (1st ed.). Elsevier. https://doi.org/10.1016/B978-0-12-815226-3.00018-1